

Asynchronous Communication: Capacity Bounds and Suboptimality of Training

Aslan Tchamkerten, Venkat Chandar, and Gregory W. Wornell *Fellow, IEEE*

Abstract—Several aspects of the problem of asynchronous point-to-point communication without feedback are developed when the source is highly intermittent. In the system model of interest, the codeword is transmitted at a random time within a prescribed window whose length corresponds to the level of asynchronism between the transmitter and the receiver. The decoder operates sequentially and communication rate is defined as the ratio between the message size and the elapsed time between when transmission commences and when the decoder makes a decision.

For such systems, general upper and lower bounds on capacity as a function of the level of asynchronism are established, and are shown to coincide in some nontrivial cases. From these bounds, several properties of this asynchronous capacity are derived. In addition, the performance of training-based schemes is investigated. It is shown that such schemes, which implement synchronization and information transmission on separate degrees of freedom in the encoding, cannot achieve the asynchronous capacity in general, and that the penalty is particularly significant in the high-rate regime.

Index Terms—asynchronous communication; bursty communication; error exponents; sequential decoding; sparse communication; synchronization

I. INTRODUCTION

INFORMATION-THEORETIC analysis of communication systems frequently ignores synchronization issues. In many applications where large amounts of data are to be transmitted, such simplifications may be justified. Simply prepending a suitable synchronization preamble to the initial data incurs negligible overhead yet ensures that the transmitter and the receiver are synchronized. In turn, various coding techniques (e.g., graph based codes, polar codes) may guarantee delay optimal communication for data transmission in the sense that they can achieve the capacity of the synchronous channel.

In quantifying the impact due to a lack of synchronization between a transmitter and a receiver, it is important to note that asynchronism is a relative notion that depends on the size of the data to be transmitted. For instance, in the above “low

asynchronism” setting it is implicitly assumed that the data is large with respect to the timing uncertainty.

In a growing number of applications, such as many involving sensor networks, data is transmitted in a bursty manner. An example would be a sensor in a monitoring system. By contrast with the previous setting, here timing uncertainty is large with respect to the data to be transmitted.

To communicate in such “high asynchronism” regimes, one can use the traditional preamble based communication scheme for each block. Alternatively, one can pursue a fundamentally different strategy in which synchronization is integrated into the encoding of the data, rather than separated from it.

To evaluate the relative merits of such diverse strategies, and more generally to explore fundamental performance limits, we recently introduced a general information-theoretic model for asynchronous communication in [3]. This model extends Shannon’s original communication model [4] to include asynchronism. In this model, the message is encoded into a codeword of fixed length, and this codeword starts being sent across a discrete memoryless channel at a time instant that is randomly and uniformly distributed over some predefined transmission window. The size of this window is known to transmitter and receiver, and the level of asynchronism in the system is governed by the size of the window with respect to the codeword length. Outside the information transmission period, whose duration equals the codeword length, the transmitter remains idle and the receiver observes noise, i.e., random output symbols. The receiver uses a sequential decoder whose scope is twofold: decide when to decode and what message to declare.

The performance measure is the communication rate which is defined as the ratio between the message size and the average delay between when transmission starts and when the message is decoded. Capacity is the supremum of achievable rates, i.e., rates for which vanishing error probability can be guaranteed in the limit of long codeword length.

The scaling between the transmission window and the codeword length that meaningfully quantifies the level of asynchronism in the system turns out to be exponential, i.e., $A = e^{\alpha n}$ where A denotes the size of the transmission window, where n denotes the codeword length, and where α denotes the asynchronism exponent. Indeed, as discussed in [3], if A scales subexponentially in n , then asynchronism doesn’t impact communication: the asynchronous capacity is equal to the capacity of the synchronous channel. By contrast, if the window size scales superexponentially, then the asynchrony is generally catastrophic. Hence, exponential asynchronism is the interesting regime and we aim to compute

This work was supported in part by an Excellence Chair Grant from the French National Research Agency (ACE project), and by the National Science Foundation under Grant No. CCF-1017772. This work was presented in part at the IEEE International Symposium on Information Theory, Toronto, Canada, July 2008 [1], and at the IEEE Information Theory Workshop, Taormina, Italy, October 2009 [2].

A. Tchamkerten is with the Department of Communications and Electronics, Telecom ParisTech, 75634 Paris Cedex 13, France. (Email: aslan.tchamkerten@telecom-paristech.fr).

V. Chandar is with MIT Lincoln Laboratory, Lexington, MA 02420 (Email: vchandar@mit.edu).

G. W. Wornell is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 (Email: gww@mit.edu).

capacity as a function of the asynchronism exponent.

For further motivation and background on the model, including a summary of related models (e.g., the insertion, deletion, and substitution channel model, and the detection and isolation model) we refer to [3, Section II]. Accordingly, we omit such material from the present paper.

The first main result in [3] is the characterization of the synchronization threshold, which is defined as the largest asynchronism exponent for which it is still possible to guarantee reliable communication—this result is recalled in Theorem 1 of Section IV.

The second main result in [3] (see [3, Theorem 1]) is a lower bound to capacity. A main consequence of this bound is that for any rate below the capacity of the synchronous channel it is possible to accommodate a non-trivial asynchronism level, i.e., a positive asynchronism exponent.

While this work focuses on rate, an alternative performance metric is the minimum energy (or, more generally, the minimum cost) needed to transmit one bit of information asynchronously. For this metric, [5], [6] establishes the capacity per unit cost for the above bursty communication setup.

We now provide a brief summary of the results contained in this paper:

- *General capacity lower bound, Theorems 2 and 1.* Theorem 2 provides a lower bound to capacity which is obtained by considering a coding scheme that performs synchronization and information transmission jointly. The derived bound results in a much simpler and often much better lower bound than the one obtained in [3, Theorem 1]. Theorem 2, which holds for arbitrary discrete memoryless channels, also holds for a natural Gaussian setting, which yields Theorem 1.
- *General capacity upper bound, Theorem 3.* This bound and the above lower bound, although not tight in general, provide interesting and surprising insights into the asynchronous capacity. For instance, Corollary 2 says that, in general, it is possible to reliably achieve a communication rate equal to the capacity of the synchronous channel while operating at a strictly positive asynchronism exponent. In other words, it is possible to accommodate both a high rate and an exponential asynchronism. Another insight is provided by Corollary 3, which relates to the very low rate communication regime. This result says that, in general, one needs to (sometimes significantly) back off from the synchronization threshold in order to be able to accommodate a positive rate. As a consequence, capacity as a function of the asynchronism exponent does not, in general, strictly increase as the latter decreases.
- *Capacity for channels with infinite synchronization threshold, Theorem 4.* For the class of channels for which there exists a particular channel input which can't be confused with noise, a closed-form expression for capacity is established.
- *Suboptimality of training based schemes, Theorem 6, Corollaries 4 and 5.* These results show that communication strategies that separate synchronization from information transmission do not achieve the asynchronous

capacity in general.

- *Good synchronous codes, Theorem 5.* This result may be of independent interest and relates to synchronous communication. It says that any codebook that achieves a nontrivial error probability contains a large subcodebook, whose rate is almost the same as the rate of the original codebook, and whose error probability decays exponentially with the blocklength with a suitable decoder. This result, which is a byproduct of our analysis, is a stronger version of [7, Corollary 1.9, p. 107] and its proof amounts to a tightening of some of the arguments in the proof of the latter.

It is worth noting that most of our proof techniques differ in some significant respects from more traditional capacity analysis for synchronous communication—for example, we make little use of Fano's inequality for converse arguments. The reason for this is that there are decoding error events specific to asynchronous communication. One such event is when the decoder, unaware of the information transmission time, declares a message before transmission even starts.

An outline of the paper is as follows. Section II summarizes some notational conventions and standard results we make use of throughout the paper. Section III describes the communication model of interest. Section IV contains our main results, and Section V is devoted to the proofs. Section VI contains some concluding remarks.

II. NOTATION AND PRELIMINARIES

In general, we reserve upper case letters for random variables (e.g., X) and lower case letters to denote their corresponding sample values (e.g., x), though as is customary, we make a variety of exceptions. Any potential confusion is generally avoided by context. In addition, we use x_i^j to denote the sequence x_i, x_{i+1}, \dots, x_j , for $i \leq j$. Moreover, when $i = 1$ we use the usual simpler notation x^n as an alternative to x_1^n . Additionally, \triangleq denotes “equality by definition.”

Events (e.g., \mathcal{E}) and sets (e.g., \mathcal{S}) are denoted using calligraphic fonts, and if \mathcal{E} represents an event, \mathcal{E}^c denotes its complement. As additional notation, $\mathbb{P}[\cdot]$ and $\mathbb{E}[\cdot]$ denote the probability and expectation of their arguments, respectively, $\|\cdot\|$ denotes the L_1 norm of its argument, $|\cdot|$ denotes absolute value if its argument is numeric, or cardinality if its argument is a set, $\lfloor \cdot \rfloor$ denotes the integer part of its argument, $a \wedge b \triangleq \min\{a, b\}$, and $x^+ \triangleq \max\{0, x\}$. Furthermore, we use \subset to denote nonstrict set inclusion, and use the Kronecker notation $\mathbb{1}(\mathcal{A})$ for the function that takes value one if the event \mathcal{A} is true and zero otherwise.

We also make use of some familiar order notation for asymptotics (see, e.g., [8, Chapter 3]). We use $o(\cdot)$ and $\omega(\cdot)$ to denote (positive or negative) quantities that grow strictly slower and strictly faster, respectively, than their arguments; e.g., $o(1)$ denotes a vanishing term and $n/\ln n = \omega(\sqrt{n})$. We also use $O(\cdot)$ and $\Omega(\cdot)$, defined analogously to $o(\cdot)$ and $\omega(\cdot)$, respectively, but without the strictness constraint. Finally, we use $\text{poly}(\cdot)$ to denote a function that does not grow or decay faster than polynomially in its argument.

We use $\mathbb{P}(\cdot)$ to denote the probability of its argument, and use $\mathcal{P}^{\mathcal{X}}$, $\mathcal{P}^{\mathcal{Y}}$, and $\mathcal{P}^{\mathcal{X}, \mathcal{Y}}$ to denote the set of distributions over

the finite alphabets \mathcal{X} , \mathcal{Y} , and $\mathcal{X} \times \mathcal{Y}$ respectively, and use $\mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ to denote the set of conditional distributions of the form $V(y|x)$ for $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

For a memoryless channel characterized by channel law $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, the probability of the output sequence $y^n \in \mathcal{Y}^n$ given an input sequence $x^n \in \mathcal{X}^n$ is

$$Q(y^n|x^n) \triangleq \prod_{i=1}^n Q(y_i|x_i).$$

Throughout the paper, Q always refers to the underlying channel and C denotes its synchronous capacity.

Additionally, we use J_X and J_Y to denote the left and right marginals, respectively, of the joint distribution $J \in \mathcal{P}^{\mathcal{X}, \mathcal{Y}}$, i.e.,

$$J_X(x) \triangleq \sum_{y \in \mathcal{Y}} J(x, y) \quad \text{and} \quad J_Y(y) \triangleq \sum_{x \in \mathcal{X}} J(x, y).$$

We define all information measures relative to the natural logarithm. Thus, the entropy associated with $P \in \mathcal{P}^{\mathcal{X}}$ is¹

$$H(P) \triangleq - \sum_{x \in \mathcal{X}} P(x) \ln P(x),$$

and the conditional entropy associated with $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ and $P \in \mathcal{P}^{\mathcal{X}}$ is

$$H(Q|P) \triangleq - \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} Q(y|x) \ln Q(y|x).$$

Similarly, the mutual information induced by $J(\cdot, \cdot) \in \mathcal{P}^{\mathcal{X}, \mathcal{Y}}$ is

$$I(J) \triangleq \sum_{(x, y) \in \mathcal{X} \times \mathcal{Y}} J(x, y) \ln \frac{J(x, y)}{J_X(x)J_Y(y)},$$

so

$$I(PQ) \triangleq \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} Q(y|x) \ln \frac{Q(y|x)}{(PQ)_Y(y)}$$

for $P \in \mathcal{P}^{\mathcal{X}}$ and $W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$. Furthermore, the information divergence (Kullback-Leibler distance) between $P_1 \in \mathcal{P}^{\mathcal{X}}$ and $P_2 \in \mathcal{P}^{\mathcal{X}}$ is

$$D(P_1 \| P_2) \triangleq \sum_{x \in \mathcal{X}} P_1(x) \ln \frac{P_1(x)}{P_2(x)},$$

and conditional information divergence is denoted using

$$\begin{aligned} D(W_1 \| W_2 | P) &\triangleq \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} W_1(y|x) \ln \frac{W_1(y|x)}{W_2(y|x)} \\ &\triangleq D(PW_1 \| PW_2), \end{aligned}$$

where $P \in \mathcal{P}^{\mathcal{X}}$ and $W_1, W_2 \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$. As a specialized notation, we use

$$D_B(\epsilon_1 \| \epsilon_2) \triangleq \epsilon_1 \ln \left(\frac{\epsilon_1}{\epsilon_2} \right) + (1 - \epsilon_1) \ln \left(\frac{1 - \epsilon_1}{1 - \epsilon_2} \right)$$

to denote the divergence between Bernoulli distributions with parameters $\epsilon_1, \epsilon_2 \in [0, 1]$.

¹In the definition of all such information measures, we use the usual convention $0 \ln(0/0) = 0$.

We make frequent use of the method of types [7, Chapter 1.2]. In particular, \hat{P}_{x^n} denotes the empirical distribution (or type) of a sequence $x^n \in \mathcal{X}^n$, i.e.,²

$$\hat{P}_{x^n}(x) \triangleq \frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i = x).$$

The joint empirical distribution $\hat{P}_{(x^n, y^n)}$ for a sequence pair (x^n, y^n) is defined analogously, i.e.,

$$\hat{P}_{x^n, y^n}(x, y) \triangleq \frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i = x, y_i = y),$$

and, in turn, a sequence y^n is said to have a conditional empirical distribution $\hat{P}_{y^n|x^n} \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ given x^n if for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\hat{P}_{x^n, y^n}(x, y) \triangleq \hat{P}_{x^n}(x) \hat{P}_{y^n|x^n}(y|x).$$

As additional notation, $P \in \mathcal{P}^{\mathcal{X}}$ is said to be an n -type if $nP(x)$ is an integer for all $x \in \mathcal{X}$. The set of all n -types over an alphabet \mathcal{X} is denoted using $\mathcal{P}_n^{\mathcal{X}}$. The n -type class of P , denoted using \mathcal{T}_P^n , is the set of all sequences x^n that have type P , i.e., such that $\hat{P}_{x^n} = P$. A set of sequences is said to have constant composition if they belong to the same type class. When clear from the context, we sometimes omit the superscript n and simply write \mathcal{T}_P . For distributions on the alphabet $\mathcal{X} \times \mathcal{Y}$ the set of joint n -types $\mathcal{P}_n^{\mathcal{X}, \mathcal{Y}}$ is defined analogously. The set of sequences y^n that have a conditional type W given x^n is denoted by $\mathcal{T}_W(x^n)$, and $\mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}$ denotes the set of empirical conditional distributions, i.e., the set of $W \in \mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}$ such that $W = \hat{P}_{y^n|x^n}(y|x)$ for some $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$.

Finally, the following three standard type results are often used in our analysis.

Fact 1 ([7, Lemma 1.2.2]):

$$\begin{aligned} |\mathcal{P}_n^{\mathcal{X}}| &\leq (n+1)^{|\mathcal{X}|} \\ |\mathcal{P}_n^{\mathcal{X}, \mathcal{Y}}| &\leq (n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|} \\ |\mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}| &\leq (n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|}. \end{aligned}$$

Fact 2 ([7, Lemma 1.2.6]): If X^n is independent and identically distributed (i.i.d.) according to $P_1 \in \mathcal{P}^{\mathcal{X}}$, then

$$\frac{1}{(n+1)^{|\mathcal{X}|}} e^{-nD(P_2 \| P_1)} \leq \mathbb{P}(X^n \in \mathcal{T}_{P_2}) \leq e^{-nD(P_2 \| P_1)},$$

for any $P_2 \in \mathcal{P}_n^{\mathcal{X}}$.

Fact 3 ([7, Lemma 1.2.6]): If the input $x^n \in \mathcal{X}^n$ to a memoryless channel $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ has type $P \in \mathcal{P}^{\mathcal{X}}$, then the probability of observing a channel output sequence Y^n which lies in $\mathcal{T}_W(x^n)$ satisfies

$$\begin{aligned} \frac{1}{(n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|}} e^{-nD(W \| Q|P)} &\leq \mathbb{P}(Y^n \in \mathcal{T}_W(x^n) | x^n) \\ &\leq e^{-nD(W \| Q|P)} \end{aligned}$$

for any $W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ such that $\mathcal{T}_W(x^n)$ is non-empty.

²When the sequence that induces the empirical type is clear from context, we omit the subscript and write simply \hat{P} .

III. MODEL AND PERFORMANCE CRITERION

The asynchronous communication model of interest captures the setting where infrequent delay-sensitive data must be reliably communicated. For a discussion of this model and its connections with related communication and statistical models we refer to [3, Section II].

We consider discrete-time communication without feedback over a discrete memoryless channel characterized by its finite input and output alphabets \mathcal{X} and \mathcal{Y} , respectively, and transition probability matrix $Q(y|x)$, for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. Without loss of generality, we assume that for all $y \in \mathcal{Y}$ there is some $x \in \mathcal{X}$ for which $Q(y|x) > 0$.

There are $M \geq 2$ messages $m \in \{1, 2, \dots, M\}$. For each message m , there is an associated codeword

$$c^n(m) \triangleq c_1(m) c_2(m) \cdots c_n(m),$$

which is a string of n symbols drawn from \mathcal{X} . The M codewords form a codebook \mathcal{C}_n (whence $|\mathcal{C}_n| = M$). Communication takes place as follows. The transmitter selects a message m randomly and uniformly over the message set and starts sending the corresponding codeword $c^n(m)$ at a random time ν , unknown to the receiver, independent of $c^n(m)$, and uniformly distributed over $\{1, 2, \dots, A\}$, where $A \triangleq e^{n\alpha}$ is referred to as the *asynchronism level* of the channel, with α termed the associated *asynchronism exponent*. The transmitter and the receiver know the integer parameter $A \geq 1$. The special case $A = 1$ (i.e., $\alpha = 0$) corresponds to the classical synchronous communication scenario.

When a codeword is transmitted, a noise-corrupted version of the codeword is obtained at the receiver. When the transmitter is silent, the receiver observes only noise. To characterize the output distribution when no input is provided to the channel, we make use of a specially designated “no-input” symbol \star in the input alphabet \mathcal{X} , as depicted in Figs. 1 and 2. Specifically,

$$Q_\star \triangleq Q(\cdot|\star) \quad (1)$$

characterizes the noise distribution of the channel. Hence, conditioned on the value of ν and on the message m to be conveyed, the receiver observes independent symbols $Y_1, Y_2, \dots, Y_{A+n-1}$ distributed as follows. If

$$t \in \{1, 2, \dots, \nu - 1\}$$

or

$$t \in [\nu + n, \nu + n + 1, \dots, A + n - 1],$$

the distribution of Y_t is Q_\star . If

$$t \in \{\nu, \nu + 1, \dots, \nu + n - 1\},$$

the distribution of Y_t is $Q(\cdot|c_{t-\nu+1}(m))$. Note that since the transmitter can choose to be silent for arbitrary portions of its length- n transmission as part of its message-encoding strategy, the symbol \star is eligible for use in the codebook design.

The decoder takes the form of a sequential test (τ, ϕ) , where τ is a stopping time, bounded by $A + n - 1$, with respect to the output sequence Y_1, Y_2, \dots , indicating when decoding happens, and where ϕ denotes a decision rule that declares the decoded message; see Fig. 2. Recall that a stopping time

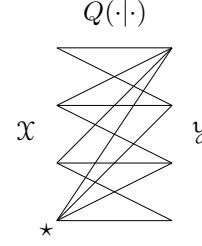


Fig. 1. Graphical depiction of the transmission matrix for an asynchronous discrete memoryless channel. The “no input” symbol \star is used to characterize the channel output when the transmitter is silent.

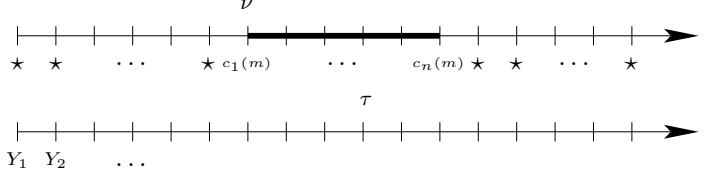


Fig. 2. Temporal representation of the channel input sequence (upper axis) and channel output sequence (lower axis). At time ν message m starts being sent and decoding occurs at time τ . Since ν is unknown at the receiver, the decoding time may be before the entire codeword has been received, potentially (but not necessarily) resulting in a decoding error.

τ (deterministic or randomized) is an integer-valued random variable with respect to a sequence of random variables $\{Y_i\}_{i=1}^\infty$ so that the event $\{\tau = t\}$, conditioned on $\{Y_i\}_{i=1}^t$, is independent of $\{Y_i\}_{i=t+1}^\infty$, for all $t \geq 1$. The function ϕ is then defined as any \mathcal{F}_τ -measurable map taking values in $\{1, 2, \dots, M\}$, where $\mathcal{F}_1, \mathcal{F}_2, \dots$ is the natural filtration induced by the process Y_1, Y_2, \dots .

A code is an encoder/decoder pair $(\mathcal{C}, (\tau, \phi))$.³

The performance of a code operating over an asynchronous channel is quantified as follows. First, we define the maximum (over messages), time-averaged decoding error probability⁴

$$\mathbb{P}(\mathcal{E}) = \max_m \frac{1}{A} \sum_{t=1}^A \mathbb{P}_{m,t}(\mathcal{E}), \quad (2)$$

where \mathcal{E} indicates the event that the decoded message does not correspond to the sent message, and where the subscripts m, t indicate the conditioning on the event that message m starts being sent at time $\nu = t$. Note that by definition we have

$$\mathbb{P}_{m,t}(\mathcal{E}) = \mathbb{P}_{m,t}(\phi(Y^\tau) \neq m).$$

Second, we define communication rate with respect to the average elapsed time between the time the codeword starts being sent and the time the decoder makes a decision, i.e.,

$$R = \frac{\ln M}{\Delta}, \quad (3)$$

where

$$\Delta = \max_m \frac{1}{A} \sum_{t=1}^A \mathbb{E}_{m,t}(\tau - t)^+, \quad (4)$$

³Note that the proposed asynchronous discrete-time communication model still assumes some degree of synchronization since transmitter and receiver are supposed to have access to clocks ticking at unison. This is sometimes referred to as frame asynchronous symbol synchronous communication.

⁴Note that there is a small abuse of notation as $\mathbb{P}(\mathcal{E})$ need not be a probability.

where x^+ denotes $\max\{0, x\}$, and where $\mathbb{E}_{m,t}$ denotes the expectation with respect to $\mathbb{P}_{m,t}$.⁵

With these definitions, the class of communication strategies of interest is as follows.

Definition 1 ((R, α) Coding Scheme): A pair (R, α) with $R \geq 0$ and $\alpha \geq 0$ is achievable if there exists a sequence $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ of codes, indexed by the codebook length n , that asymptotically achieves a rate R at an asynchronism exponent α . This means that for any $\epsilon > 0$ and every n large enough, the code $(\mathcal{C}_n, (\tau_n, \phi_n))$

- 1) operates under asynchronism level $A_n = e^{(\alpha - \epsilon)n}$;
- 2) yields a rate at least equal to $R - \epsilon$;
- 3) achieves a maximum error probability of at most ϵ .

An (R, α) coding scheme is a sequence $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ that achieves the rate-exponent pair (R, α) .

In turn, capacity for our model is defined as follows.

Definition 2 (Asynchronous Capacity): For given $\alpha \geq 0$, the asynchronous capacity $R(\alpha)$ is the supremum of the set of rates that are achievable at asynchronism exponent α . Equivalently, the asynchronous capacity is characterized by $\alpha(R)$, defined as the supremum of the set of asynchronism exponents that are achievable at rate $R \geq 0$.

Accordingly, we use the term ‘‘asynchronous capacity’’ to designate either $R(\alpha)$ or $\alpha(R)$. While $R(\alpha)$ may have the more natural immediate interpretation, most of our results are more conveniently expressed in terms of $\alpha(R)$.

In agreement with our notational convention, the capacity of the synchronous channel, which corresponds to the case where $\alpha = 0$, is simply denoted by C instead of $R(0)$. Throughout the paper we only consider channels with $C > 0$.

Remark 1: One could alternatively consider the rate with respect to the duration the transmitter occupies the channel and define it with respect to the block length n . In this case capacity is a special case of the general asynchronous capacity per unit cost result [5, Theorem 1].

In [3], [9] it is shown that reliable communication is possible if and only if the asynchronism exponent α does not exceed a limit referred to as the ‘‘synchronization threshold.’’

Theorem 1 ([3, Theorem 2], [9]): If the asynchronism exponent is strictly smaller than the *synchronization threshold*

$$\alpha_o \triangleq \max_x D(Q(\cdot|x) \| Q_\star) = \alpha(R = 0),$$

then there exists a coding scheme $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ that achieves a maximum error probability tending to zero as $n \rightarrow \infty$.

Conversely, any coding scheme $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ that operates at an asynchronism exponent strictly greater than the synchronization threshold, achieves (as $n \rightarrow \infty$) a maximum probability of error equal to one.

Moreover,⁶

$$\alpha_o > 0 \quad \text{if and only if} \quad C > 0.$$

A few comments are in order. The cause of unreliable communication above the synchronization threshold is the following. When asynchronism is so large, with probability

approaching one pure noise mimics a codeword for *any* codebook (regardless of the rate) before the actual codeword even starts being sent.⁷ This results in an error probability of at least $1/2$ since, by our model assumption, the message set contains at least two messages. On the other hand, below the synchronization threshold reliable communication is possible. If the codebook is properly chosen, the noise won’t mimic any codeword with probability tending to one, which allows the decoder to reliably detect the sent message.

Note that

$$\alpha_o = \infty$$

if and only if pure noise can’t generate all channel outputs, i.e., if and only if $Q_\star(y) = 0$ for some $y \in \mathcal{Y}$. Indeed, in this case it is possible to avoid the previously mentioned decoding confusion by designing codewords (partly) composed of symbols that generate channel outputs which are impossible to generate with pure noise.

The last claim in Theorem 1 says that reliable asynchronous communication is possible if and only if reliable synchronous communication is possible. That the former implies the latter is obvious since asynchronism can only hurt communication. That the latter implies the former is perhaps less obvious, and a high-level justification is as follows. When $C > 0$, at least two channel inputs yield different conditional output distributions, for otherwise the input-output mutual information is zero regardless of the input distribution. Hence, $Q(\cdot|\star) \neq Q(\cdot|x)$ for some $x \neq \star$. Now, by designing codewords mainly composed of x it is possible to reliably signal the codeword’s location to the decoder even under an exponential asynchronism, since the channel outputs look statistically different than noise during the message transmission. Moreover, if the message set is small enough, it is possible to guarantee reliable message location and successfully identify which message from the message set was sent. Therefore, exponential asynchronism can be accommodated, hence $\alpha_o > 0$.

Finally, it should be pointed out that in [3] all the results are stated with respect to average (over messages) delay and error probability in place of maximum (over messages) delay and error probability as in this paper. Nevertheless, the same results hold in the latter case as discussed briefly later at the end of Section V.

IV. MAIN RESULTS

This section is divided into two parts. In Section IV-A, we provide general upper and lower bounds on capacity, and derive several of its properties. In Section IV-B, we investigate the performance limits of training-based schemes and establish their suboptimality in a certain communication regime. Since both sections can be read independently, the practically inclined reader may read Section IV-B first.

All of our results assume a uniform distribution on ν . Nevertheless, this assumption is not critical in our proofs. The results can be extended to non-uniform distributions by following the same arguments as those used to establish

⁵Note that $\mathbb{E}_{m,t}(\tau_n - t)^+$ should be interpreted as $\mathbb{E}_{m,t}((\tau_n - t)^+)$.

⁶This claim appeared in [3, p. 4515].

⁷This follows from the converse of [9, Theorem], which says that above α_o , even the codeword of a single codeword codebook is mislocated with probability tending to one.

asynchronous capacity per unit cost for non-uniform ν [5, Theorem 5].

A. General Bounds on Asynchronous Capacity

A decoder at the output of an asynchronous channel should discriminate between hypothesis “noise” and hypothesis “message,” which correspond to the situations when the transmitter is idle and when it transmits a codeword, respectively. Intuitively, the more these hypotheses are statistically far apart—by means of an appropriate codebook design—the larger the level of asynchronism which can be accommodated for a given communication rate.

More specifically, a code should serve the dual purpose of minimizing the “false-alarm” and “miss” error probabilities.

False-alarm refers to the event where the decoder outputs a message before a message is sent. As such, this event contributes to lower the rate—since it is defined with respect to the receiver’s decoding delay $\mathbb{E}(\tau - \nu)^+$ —at the expense of the error probability. As an extreme case, by immediately decoding, *i.e.*, by setting $\tau = 1$, we get an infinite rate and an error probability (asymptotically) equal to one. As it turns out, the false-alarm probability should be exponentially small to allow reliable communication under exponential asynchronism.

The miss event refers to the scenario where the decoder fails to recognize the sent message during transmission, *i.e.*, the message output looks like it was generated by noise. This event impacts the rate and, to a smaller extent, also the error probability. In fact, when the sent message is missed, the reaction delay is usually huge, of the order of A . Therefore, to guarantee a positive rate under exponential asynchronism the miss error probability should also be exponentially small.

Theorem 2 below provides a lower bound on the asynchronous capacity. The proof of this theorem is obtained by analyzing a coding scheme which performs synchronization and information transmission jointly. The codebook is a standard i.i.d. random code across time and messages and its performance is governed by the Chernoff error exponents for discriminating hypothesis “noise” from hypothesis “message.”

Theorem 2 (Lower Bound on Asynchronous Capacity):

Let $\alpha \geq 0$ and let $P \in \mathcal{P}^{\mathcal{X}}$ be some input distribution such that at least one of the following inequalities

$$\begin{aligned} D(V \parallel (PQ)_y) &\geq \alpha \\ D(V \parallel Q_*) &\geq \alpha \end{aligned}$$

holds for all distributions $V \in \mathcal{P}^{\mathcal{Y}}$, *i.e.*,

$$\min_{V \in \mathcal{P}^{\mathcal{Y}}} \max\{D(V \parallel (PQ)_y), D(V \parallel Q_*)\} \geq \alpha.$$

Then, the rate-exponent pair $(R = I(PQ), \alpha)$ is achievable. Thus, maximizing over all possible input distributions, we have the following lower bound on $\alpha(R)$ in Definition 2:

$$\alpha(R) \geq \alpha_-(R) \quad R \in (0, C] \quad (5)$$

where

$$\alpha_-(R) \triangleq \max_{\substack{P \in \mathcal{P}^{\mathcal{X}}: \\ I(PQ) \geq R}} \min_{V \in \mathcal{P}^{\mathcal{Y}}} \max\{D(V \parallel (PQ)_y), D(V \parallel Q_*)\}. \quad (6)$$

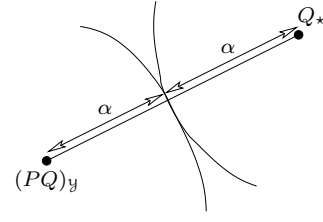


Fig. 3. If α is at most the “half-distance” between distributions $(PQ)_y$ and Q_* , then (α, R) with $R = I(PQ)$ is achievable.

The analysis of the coding scheme that yields Theorem 2 is actually tight in the sense that the coding scheme achieves (6) with equality (see proof of Theorem 2 and remark p. 14.)

Theorem 2 provides a simple explicit lower bound on capacity. The distribution $(PQ)_y$ corresponds to the channel output when the input to the channel is distributed according to P . The asynchronism exponent that can be accommodated for given P and Q_* can be interpreted as being the “equidistant point” between distributions $(PQ)_y$ and Q_* , as depicted in Fig. 3. Maximizing over P such that $I(PQ) \geq R$ gives the largest such exponent that can be achieved for rate R communication.

Note that (6) is much simpler to evaluate than the lower bound given by [3, Theorem 2]. Moreover, the former is usually a better bound than the latter and it exhibits an interesting feature of $\alpha(R)$ in the high rate regime. This feature is illustrated in Example 1 to come.

Theorem 2 extends to the following continuous alphabet Gaussian setting:

Corollary 1 (Asynchronous Gaussian channel): Suppose that for a real input x the decoder receives $Y = x + Z$, where $Z \sim \mathcal{N}(0, 1)$. When there is no input to the channel, $Y = Z$, so $Q_* = \mathcal{N}(0, 1)$. The input is power constrained so that all codewords $c^n(m)$ must satisfy $\frac{1}{n} \sum_{i=1}^n c_i(m)^2 \leq p$ for a given constant $p > 0$. For this channel we have

$$\alpha(R) \geq \max_{\substack{P: I(PQ) \geq R \\ \mathbb{E}_P X^2 \leq p}} \min_V \max\{D(V \parallel (PQ)_y), D(V \parallel Q_*)\}, \quad (7)$$

for $R \in (0, C]$ where P and V in the optimization are distributions over the reals.

If we restrict the outer maximization in (7) to be over Gaussian distributions only, it can be shown that the best input has a mean μ that is as large as possible, given the rate and power constraints. More precisely, μ and R satisfy

$$R = \frac{1}{2} \ln(1 + p - \mu^2),$$

and the variance of the optimal Gaussian input is $p - \mu^2$. The intuition for choosing such parameters is that a large mean helps the decoder to distinguish the codeword from noise—since the latter has a mean equal to zero. What limits the mean is both the power constraint and the variance needed to ensure sufficient mutual information to support communication at rate R .

Proof of Corollary 1: The proof uses a standard quantization argument similar to that in [10], and therefore we provide only a sketch of the proof. From the given the continuous time Gaussian channel, we can form a discrete alphabet channel for which we can apply Theorem 2.

More specifically, pick a discrete input distribution P that satisfies the power constraint. The output is discretized within $[-L/2, L/2]$ into constant size Δ intervals so that $L \rightarrow \infty$ as $\Delta \rightarrow 0$. The output of the quantized channel corresponds to the mid-value of the interval which contains the output of the Gaussian channel. If the output of the Gaussian channel falls below $-L/2$, the quantized value is set to be $-L/2$, and if the output of the Gaussian channel falls above $L/2$, the quantized value is set to be $L/2$.

For each quantized channel we apply Theorem 2, then let delta tend to zero. One can then verify that the achieve bound corresponds to (7), which shows that Theorem 2 also holds for the continuous alphabet Gaussian setting of Theorem 1. ■

The next result provides an upper bound to the asynchronous capacity for channels with finite synchronization threshold—see Theorem 1:

Theorem 3 (Upper Bound on Asynchronous Capacity):

For any channel Q such that $\alpha_o < \infty$, and any $R > 0$, we have that

$$\alpha(R) \leq \max_{\mathcal{S}} \min\{\alpha_1, \alpha_2\} \triangleq \alpha_+(R), \quad (8)$$

where

$$\alpha_1 \triangleq \delta(I(P_1 Q) - R + D((P_1 Q)_y \| Q_*)) \quad (9)$$

$$\alpha_2 \triangleq \min_{W \in \mathcal{P}^{y|x}} \max\{D(W \| Q | P_2), D(W \| Q_* | P_2)\} \quad (10)$$

with

$$\mathcal{S} \triangleq \left\{ (P_1, P_2, P'_1, \delta) \in (\mathcal{P}^x)^3 \times [0, 1] : \right. \\ \left. I(P_1 Q) \geq R, P_2 = \delta P_1 + (1 - \delta) P'_1 \right\}. \quad (11)$$

If $\alpha_o = \infty$, then

$$\alpha(R) \leq \max_{P_2} \alpha_2 \quad (12)$$

for $R \in (0, C]$.

The terms α_1 and α_2 in (8) reflect the false-alarm and miss constraints alluded to above (see discussion before Theorem 2). If $\alpha > \alpha_1$, then with high probability the noise will mimic a message before transmission starts. Instead, if $\alpha > \alpha_2$ then reliable communication at a positive rate is impossible since no code can guarantee a sufficiently low probability of missing the sent codeword.

The parameter δ in (9) and (11) essentially represents the ratio between the reaction delay $\mathbb{E}(\tau - \nu)^+$ and the blocklength—which need not coincide. Loosely speaking, for a given asynchronism level a smaller δ , or, equivalently, a smaller $\mathbb{E}(\tau - \nu)^+$, increases the communication rate at the expense of a higher false-alarm error probability. The intuition for this is that a decoder that achieves a smaller reaction delay sees, on average, “fewer” channel outputs before stopping. As a consequence, the noise is more likely to lead such a decoder into confusion. A similar tension arises between

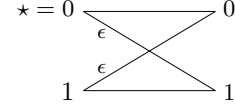


Fig. 4. A channel for which $\alpha(R)$ is discontinuous at $R = C$.

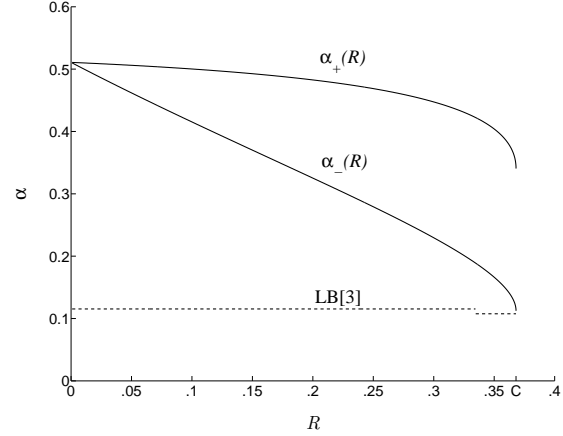


Fig. 5. Capacity upper and lower bounds on the asynchronous capacity of the channel of Fig. 4 with $\epsilon = 0.1$ and $\star = 0$. $\alpha_-(R)$ represents the lower bound given by Theorem 2, LB[3] represents the lower bound obtained in [3, Theorem 1], and $\alpha_+(R)$ represents the upper bound given by Theorem 3.

communication rate and the miss error probability. The optimization over the set \mathcal{S} attempts to strike the optimal tradeoff between the communication rate, the false-alarm and miss error probabilities, as well as the reaction delay as a fraction of the codeword length.

For channels with infinite synchronization threshold, Theorem 4 to come establishes that the bound given by (12) is actually tight.

The following examples provide some useful insights.

Example 1: Consider the binary symmetric channel depicted in Fig. 4, which has the property that when no input is supplied to the channel, the output distribution is asymmetric. For this channel, in Fig. 5 we plot the lower bound on $\alpha(R)$ given by (6) (curve $\alpha_-(R)$) and the lower bound given by [3, Theorem 1] (the dashed line LB[3]).⁸ The $\alpha_+(R)$ curve correspond to the upper bound on $\alpha(R)$ given by Theorem 3. For these plots, the channel parameter is $\epsilon = 0.1$.

The discontinuity of $\alpha(R)$ at $R = C$ (since $\alpha(R)$ is clearly equal to zero for $R > C$) implies that we do not need to back off from the synchronous capacity in order to operate under

⁸Due to the complexity of evaluating the lower bound given by [3, Theorem 1], the curves labeled LB[3] are actually upper bounds on this lower bound. We believe these bounds are fairly tight, but in any case we see that the resulting upper bounds are below the lower bounds given by (6).

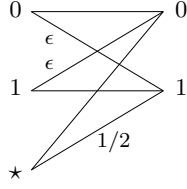


Fig. 6. Channel for which $\alpha(R)$ is continuous at $R = C$.

exponential asynchronism.⁹

Note next that the $\alpha_-(R)$ is better than LB[3] for all rates. In fact, empirical evidence suggests that $\alpha_-(R)$ is better than LB[3] in general. Additionally, note that $\alpha_-(R)$ and $\alpha_+(R)$ are not tight.

Next, we show how another binary symmetric channel has some rather different properties.

Example 2: Consider the binary symmetric channel depicted in Fig. 6, which has the property that when no input is provided to the channel the output distribution is symmetric. When used synchronously, this channel and that of Example 1 are completely equivalent, regardless of the crossover probability ϵ . Indeed, since the \star input symbol in Fig. 6 produces 0 and 1 equiprobably, this input can be ignored for coding purposes and any code for this channel achieves the same performance on the channel in Fig. 4.

However, this equivalence no longer holds when the channels are used asynchronously. To see this, we plot the corresponding upper and lower bounds on performance for this channel in Fig. 7. Comparing curve $\alpha_-(R)$ in Fig. 5 with curve $\alpha_+(R)$ in Fig. 7, we see that asynchronous capacity for the channel of Fig. 4 is always larger than that of the current example. Moreover, since there is no discontinuity in exponent at $R = C$ in our current example, the difference is pronounced at $R = C = 0.368\dots$; for the channel of Fig. 4 we have $\alpha(C) \approx 0.12 > 0$.

The discontinuity of $\alpha(R)$ at $R = C$ observed in Example 1 is in fact typical, holding in all but one special case.

Corollary 2 (Discontinuity of $\alpha(R)$ at $R = C$): We have $\alpha(C) = 0$ if and only if Q_\star corresponds to the (unique) capacity-achieving output distribution of the synchronous channel.

By Corollary 2, for the binary symmetric channel of Example 1, $\alpha(R)$ is discontinuous at $R = C$ whenever $\epsilon \neq 1/2$. To see this, note that the capacity achieving output distribution of the synchronous channel assigns equal weights to \star and 1, differently than Q_\star .

The justification for the discontinuity in Example 1 is as follows. Since the capacity-achieving output distribution of the synchronous channel (Bernoulli(1/2)) is “biased” with

⁹To have a better sense of what it means to be able to decode under exponential asynchronism and, more specifically, at $R = C$, consider the following numerical example. Consider a codeword length n equal to 150. Then $\alpha = .12$ yields asynchronism level $A = e^{n\alpha} \approx 6.5 \times 10^7$. If the codeword is, say, 30 centimeters long, then this means that the decoder can reliably sequentially decode the sent message, with minimal delay (were the decoder cognizant of ν , it couldn’t achieve a smaller decoding delay since we operate at the synchronous capacity), within 130 kilometers of mostly noisy data!

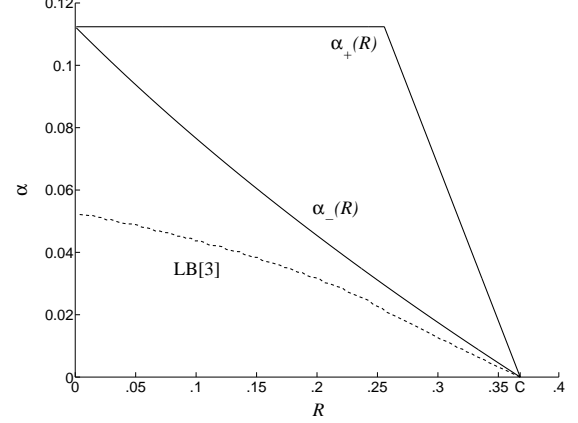


Fig. 7. Capacity upper and lower bounds on the asynchronous capacity of the channel of Fig. 6 with $\epsilon = 0.1$. $\alpha_-(R)$ represents the lower bound given by Theorem 2, LB[3] represents the lower bound obtained in [3, Theorem 1], and $\alpha_+(R)$ represents the upper bound given by Theorem 3.

respect to the noise distribution Q_\star , hypothesis “message” and “noise” can be discriminated with exponentially small error probabilities. This, in turn, enables reliable detection of the sent message under exponential asynchronism. By contrast, for the channel of Example 2, $\alpha(R)$ is continuous at $R = C$, regardless of ϵ .

Proof of Corollary 2: From Theorem 2, a strictly positive asynchronism exponent can be achieved at $R = C$ if Q_\star differs from the synchronous capacity-achieving output distribution—(6) is strictly positive for $R = C$ whenever Q_\star differs from the synchronous capacity-achieving output distribution since the divergence between two distributions is zero only if they are equal.

Conversely, suppose Q_\star is equal to the capacity-achieving output distribution of the synchronous channel. We show that for any (R, α) coding scheme where $R = C$, α is necessarily equal to zero.

From Theorem 3,

$$\alpha(R) \leq \max_{\mathcal{S}} \alpha_1$$

where \mathcal{S} and α_1 are given by (11) and (9), respectively. Since $R = C$, $I(P_1 Q) = C$, and since $Q_\star = (P_1 Q)_y$, we have $D((P_1 Q)_y || Q_\star) = 0$. Therefore, $\alpha_1 = 0$ for any δ , and we conclude that $\alpha(C) = 0$. ■

In addition to the discontinuity at $R = C$, $\alpha(R)$ may also be discontinuous at rate zero:

Corollary 3 (Discontinuity of $\alpha(R)$ at $R = 0$): If

$$\alpha_o > \max_{x \in \mathcal{X}} D(Q_\star || Q(\cdot | x)), \quad (13)$$

then $\alpha(R)$ is discontinuous at rate $R = 0$.

Example 3: Channels that satisfy (13) include those for which the following two conditions hold: \star can’t produce all channel outputs, and if a channel output can be produced by \star , then it can also be produced by any other input symbol. For these channels (13) holds trivially; the right-hand side term is finite and the left-hand side term is infinite. The simplest such channel is the Z-channel depicted in Fig. 8 with $\epsilon \in (0, 1)$.

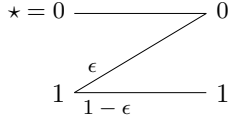


Fig. 8. Channel for which $\alpha(R)$ is discontinuous at $R = 0$, assuming $\epsilon \in (0, 1)$.

Note that if $\epsilon = 0$, (13) doesn't hold since both the left-hand side term and the right-hand side term are infinite. In fact, if $\epsilon = 0$ then asynchronism doesn't impact communication; rates up to the synchronous capacity can be achieved regardless of the level of asynchronism, i.e.,

$$\alpha(R) = \alpha_o = \infty \quad R \in [0, C].$$

To see this, note that by prepending a 1 to each codeword suffices to guarantee perfect synchronization without impacting rate (asymptotically).

More generally, asynchronous capacity for channels with infinite synchronization threshold is established in Theorem 4 to come.

An intuitive justification for the possible discontinuity of $\alpha(R)$ at $R = 0$ is as follows. Consider a channel where \star cannot produce all channel outputs (such as that depicted in Fig. 8). A natural encoding strategy is to start codewords with a common preamble whose possible channel outputs differ from the set of symbols that can be generated by \star . The remaining parts of the codewords are chosen to form, for instance, a good code for the synchronous channel. Whenever the decoder observes symbols that cannot be produced by noise (a clear sign of the preamble's presence), it stops and decodes the upcoming symbols. For this strategy, the probability of decoding before the message is actually sent is clearly zero. Also, the probability of wrong message isolation conditioned on correct preamble location can be made negligible by taking codewords long enough. Similarly, the probability of missing the preamble can be made negligible by using a long enough preamble. Thus, the error probability of this training-based scheme can be made negligible, regardless of the asynchronism level.

The problem arises when we add a positive rate constraint, which translates into a delay constraint. Conditioned on missing the preamble, it can be shown that the delay $(\tau - \nu)^+$ is large, in fact of order A . It can be shown that if (13) holds, the probability of missing the preamble is larger than $1/A$. Therefore, a positive rate puts a limit on the maximum asynchronism level for which reliable communication can be guaranteed, and this limit can be smaller than α_o .

We note that it is an open question whether or not $\alpha(R)$ may be discontinuous at $R = 0$ for channels that do not satisfy (13).

Theorem 4 provides an exact characterization of capacity for the class of channels with infinite synchronization threshold, i.e., whose noise distribution Q_\star cannot produce all possible channel outputs.

Theorem 4 (Capacity when $\alpha_o = \infty$): If $\alpha_o = \infty$, then

$$\alpha(R) = \bar{\alpha} \quad (14)$$

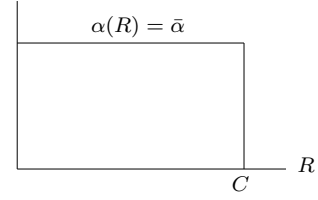


Fig. 9. Typical shape of the capacity of an asynchronous channel Q for which $\alpha_o = \infty$.

for $R \in (0, C]$, where

$$\bar{\alpha} \triangleq \max_{P \in \mathcal{P}^{\mathcal{X}}} \min_{W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}} \max\{D(W\|Q|P), D(W\|Q_\star|P)\}.$$

Therefore, when $\alpha_o = \infty$, $\alpha(R)$ is actually a constant that does not depend on the rate, as Fig. 9 depicts. Phrased differently, $R(\alpha) = C$ up to $\alpha = \bar{\alpha}$. For $\alpha > \bar{\alpha}$ we have $R(\alpha) = 0$.

Note that when $\alpha_o = \infty$, $\alpha(R)$ can be discontinuous at $R = 0$ since the right-hand side of (14) is upper bounded by

$$\max_{x \in \mathcal{X}} D(Q_\star \| Q(\cdot|x)),$$

which can be finite.¹⁰

We conclude this section with a result of independent interest related to synchronous communication, and which is obtained as a byproduct of the analysis used to prove Theorem 3. This result essentially says that any nontrivial fixed length codebook, i.e., that achieves a nontrivial error probability, contains a very good large (constant composition) sub-codebook, in the sense that its rate is almost the same as the original code, but its error probability decays exponentially with a suitable decoder. In the following theorem (\mathcal{C}_n, ϕ_n) denotes a standard code for a synchronous channel Q , with fixed length n codewords and decoding happening at time n .

Theorem 5: Fix a channel $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, let $q \in (0, 1/2)$, and let $\epsilon, \gamma > 0$ be such that $\epsilon + \gamma \in (0, l)$ with $l \in (0, 1)$. If (\mathcal{C}_n, ϕ_n) is a code that achieves an error probability ϵ , then there exists an $n_o(l, \gamma, q, |\mathcal{X}|, |\mathcal{Y}|)$ such that for all $n \geq n_o$ there exists $(\mathcal{C}'_n, \phi'_n)$ such that¹¹

- 1) $\mathcal{C}'_n \subset \mathcal{C}_n$, \mathcal{C}'_n is constant composition;
- 2) the maximum error probability is less than ϵ_n where

$$\epsilon_n = 2(n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n^{2q}/(2 \ln 2));$$

- 3) $\frac{\ln |\mathcal{C}'_n|}{n} \geq \frac{\ln |\mathcal{C}_n|}{n} - \gamma$.

Theorem 5 is a stronger version of [7, Corollary 1.9, p. 107] and its proof amounts to a tightening of some of the arguments in the proof of the latter, but otherwise follows it closely.

B. Training-Based Schemes

Practical solutions to asynchronous communication usually separate synchronization from information transmission. We investigate a very general class of such “training-based schemes” in which codewords are composed of two parts:

¹⁰To see this choose $W = Q_\star$ in the minimization (14).

¹¹We use $n_o(q)$ to denote some threshold index which could be explicitly given as a function of q .

a preamble that is common to all codewords, followed by information symbols. The decoder first attempts to detect the preamble, then decodes the information symbols. The results in this section show that such schemes are suboptimal at least in certain communication regimes. This leads to the conclusion that the separation of synchronization and information transmission is in general not optimal.

We start by defining a general class of training-based schemes:

Definition 3 (Training-Based Scheme): A coding scheme $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ is said to be *training-based* if for some $\eta \in [0, 1]$ and all n large enough

- 1) there is a common preamble across codewords of size ηn ;
- 2) the decoding time τ_n is such that the event

$$\{\tau_n = t\},$$

conditioned on the ηn observations $Y_{t-n+1}^{t-n+\eta n}$, is independent of all other observations (i.e., Y_1^{t-n} and $Y_{t-n+\eta n+1}^{A+n-1}$).

Note that Definition 3 is in fact very general. The only restrictions are that the codewords all start with the same training sequence, and that the decoder's decision to stop at any particular time should be based on the processing of (at most) ηn past output symbols corresponding to the length of the preamble.

In the sequel we use $\alpha^T(R)$ to denote the asynchronous capacity restricted to training based schemes.

Theorem 6 (Training-based scheme capacity bounds): Capacity restricted to training based schemes satisfies

$$\alpha_-^T(R) \leq \alpha^T(R) \leq \alpha_+^T(R) \quad R \in (0, C] \quad (15)$$

where

$$\begin{aligned} \alpha_-^T(R) &\triangleq m_1 \left(1 - \frac{R}{C}\right) \\ \alpha_+^T(R) &\triangleq \min \left\{ m_2 \left(1 - \frac{R}{C}\right), \alpha_+(R) \right\}, \end{aligned}$$

where the constants m_1 and m_2 are defined as

$$\begin{aligned} m_1 &\triangleq \max_{P \in \mathcal{P}^X} \min_{W \in \mathcal{P}^{Y|X}} \max\{D(W||Q|P), D(W||Q_\star|P)\} \\ m_2 &\triangleq -\ln(\min_{y \in \mathcal{Y}} Q_\star(y)), \end{aligned}$$

and where $\alpha_+(R)$ is defined in Theorem 3.

Moreover, a rate $R \in [0, C]$ training-based scheme allocates at most a fraction

$$\eta = \left(1 - \frac{R}{C}\right)$$

to the preamble.

Since $m_2 < \infty$ if and only if $\alpha_o < \infty$, the upper-bound in (15) implies:

Corollary 4 (Asynchronism in the high rate regime): For training-based schemes

$$\alpha^T(R) \xrightarrow{R \rightarrow C} 0$$

whenever $\alpha_o < \infty$.

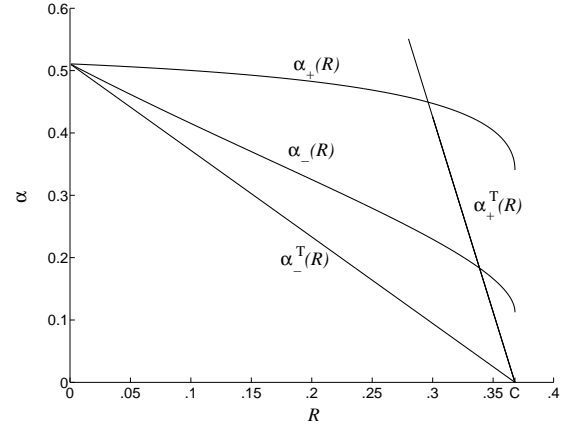


Fig. 10. Upper and lower bounds to capacity restricted to training-based schemes (TUB and TLB, respectively) for the binary symmetric channel depicted in Fig. 4 with $\epsilon = 0.1$. $\alpha_+(R)$ and $\alpha_-(R)$ represent the capacity general upper and lower bounds given by Theorems 2 and 3.

In general, $\alpha(C) > 0$ as we saw in Corollary 2. Hence a direct consequence of Corollaries 2 and 4 is that training-based schemes are suboptimal in the high rate regime. Specifically, we have the following result.

Corollary 5 (Suboptimality of training-based schemes): There exists a channel-dependent threshold R_* such that for all $R > R_*$,

$$\alpha^T(R) < \alpha(R)$$

except possibly when Q_\star corresponds to the capacity-achieving output distribution of the synchronous channel, or when the channel is degenerate, i.e., when $\alpha_o = \infty$.

The last claim of Theorem 6 says that the size of the preamble decreases (linearly) as the rate increases. This, in turn, implies that $\alpha^T(R)$ tends to zero as R approaches C . Hence, in the high rate regime most of the symbols should carry information, and the decoder should try to detect these symbols as part of the decoding process. In other words, synchronization and information transmission should be jointly performed; transmitted bits should carry information while also helping the decoder to locate the sent codeword.

If we are willing to reduce the rate, are training-based schemes still suboptimal? We do not have a definite answer to this question, but the following examples provide some insights.

Example 4: Consider the channel depicted in Fig. 4 with $\epsilon = 0.1$. In Fig. 10, we plot the upper and lower bounds to capacity restricted to training-based schemes given by Theorem 6. TLB represents the lower bound in (15) and TUB represents the $m_2(1 - R/C)$ term in the upper bound (15). $\alpha_-(R)$ and $\alpha_+(R)$ represent the general lower and upper bounds to capacity given by Theorems 2 and 3; see Fig. 5.

By comparing $\alpha_-(R)$ with TUB in Fig. 10 we observe that for rates above roughly 92% of the synchronous capacity C , training-based schemes are suboptimal.

For this channel, we observe that $\alpha_-(R)$ is always above TLB. This feature does not generalize to arbitrary crossover probabilities ϵ . Indeed, consider the channel in Fig. 4, but with

an arbitrary crossover probability ϵ , and let r be an arbitrary constant such that $0 < r < 1$. From Theorem 6, training-based schemes can achieve rate asynchronism pairs (R, α) that satisfy

$$\alpha \geq m_1(1 - R/C(\epsilon)) \quad R \in (0, C(\epsilon)].$$

For the channel at hand

$$m_1 = D_B(1/2||\epsilon),$$

hence α tends to infinity as $\epsilon \rightarrow 0$, for any fixed $R \in (0, r)$ —note that $C(\epsilon) \rightarrow 1$ as $\epsilon \rightarrow 0$.

Now, consider the random coding scheme that yields Theorem 2. This scheme, which performs synchronization and information transmission jointly, achieves for any given rate $R \in [0, C]$ asynchronism exponent (see comment after Theorem 2)

$$\alpha = \max_{\{P \in \mathcal{P}^x : I(PQ) \geq R\}} \min_{V \in \mathcal{P}^y} \max\{D(V||PQ)_y, D(V||Q_*)\}.$$

This expression is upper-bounded by¹²

$$\max_{P \in \mathcal{P}^x : I(PQ) \geq R} D(Q_*||PQ)_y, \quad (16)$$

which is bounded in the limit $\epsilon \rightarrow 0$ as long as $R > 0$.¹³ Therefore the joint synchronization-information transmission code yielding Theorem 2 can be outperformed by training-based schemes at moderate to low rate, even when the output distribution when no input is supplied is asymmetric. This shows that the general lower bound given by Theorem 2 is loose in general.

Example 5: For the channel depicted in Fig. 6 with $\epsilon = 0.1$, in Fig. 11 we plot the upper and lower bounds on capacity restricted to training-based schemes, as given by Theorem 6. For this channel it turns out that the training-based scheme upper bound $m_2(1 - R/C)$ (see Theorem 6) is loose and hence TUB = $\alpha_+(R)$ for all rates. In contrast with the example of Fig. 10, here the general lower bound $\alpha_-(R)$ is below the lower bound for the best training best schemes (TLB line).

V. ANALYSIS

In this section, we establish the theorems of Section IV.

A. Proof of Theorem 2

Let $\alpha \geq 0$ and $P \in \mathcal{P}^x$ satisfy the assumption of the theorem, i.e., be such that at least one of the following inequalities holds

$$\begin{aligned} D(V||PQ)_y &\geq \alpha \\ D(V||Q_*) &\geq \alpha \end{aligned} \quad (17)$$

for all distributions $V \in \mathcal{P}^y$, and let $A_n = e^{n(\alpha - \epsilon)}$.

The proof is based on a random coding argument associated with the following communication strategy. The codebook

¹²To see this, choose $V = Q_*$ in the minimization.

¹³Let $P^* = P^*(Q)$ be an input distribution P that maximizes (16) for a given channel. Since $R \leq I(P^*Q) \leq H(P^*)$, P^* is uniformly bounded away from 0 and 1 for all $\epsilon \geq 0$. This implies that (16) is bounded in the limit $\epsilon \rightarrow 0$.

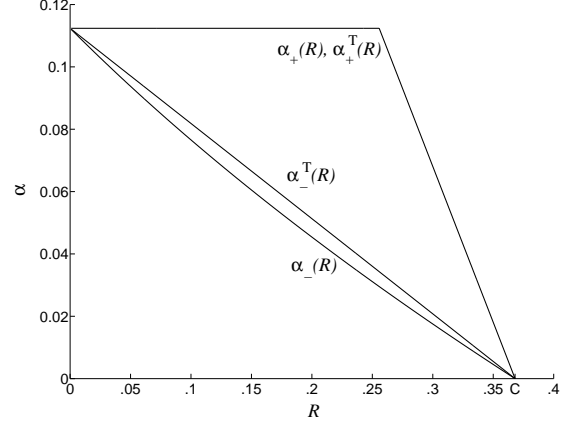


Fig. 11. Lower bound (TLB) to capacity restricted to training-based schemes for the channel of Fig. 6. $\alpha_+(R)$ and $\alpha_-(R)$ represent the capacity general upper and lower bounds given by Theorems 2 and 3. For this channel the training upper bound (TUB) coincides with $\alpha_+(R)$, and hence is not plotted separately.

$\mathcal{C} = \{c^n(m)\}_{m=1}^M$ is randomly generated so that all $c_i(m)$, $i \in \{1, 2, \dots, n\}$, $m \in \{1, 2, \dots, M\}$, are i.i.d. according to P . The sequential decoder operates according to a two-step procedure. The first step consists in making a coarse estimate of the location of the sent codeword. Specifically, at time t the decoder tries to determine whether the last n output symbols are generated by noise or by some codeword on the basis of their empirical distribution $\hat{P} = \hat{P}_{y_{t-n+1}^t}$. If $D(\hat{P}||Q_*) < \alpha$, \hat{P} is declared a “noise type,” the decoder moves to time $t+1$, and repeats the procedure, i.e., tests whether $\hat{P}_{y_{t-n+2}^{t+1}}$ is a noise type. If, instead, $D(\hat{P}||Q_*) \geq \alpha$, the decoder marks the current time as the beginning of the “decoding window,” and proceeds to the second step of the decoding procedure.

The second step consists in exactly locating and identifying the sent codeword. Once the beginning of the decoding window has been marked, the decoder makes a decision the first time that the previous n symbols are jointly typical with one of the codewords. If no such time is found within n successive time steps, the decoder stops and declares a random message. The typicality decoder operates as follows.¹⁴ Let P_m be the probability measure induced by codeword $c^n(m)$ and the channel, i.e.,

$$P_m(a, b) \triangleq \hat{P}_{c^n(m)}(a)Q(b|a) \quad (a, b) \in \mathcal{X} \times \mathcal{Y}. \quad (18)$$

At time t , the decoder computes the empirical distributions \hat{P}_m induced by $c^n(m)$ and the n output symbols y_{t-n+1}^t for all $m \in \{1, 2, \dots, M\}$. If

$$|\hat{P}_{c^n(m), y_{t-n+1}^t}(a, b) - P_m(a, b)| \leq \mu$$

for all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ and a unique index m , the decoder declares message m as the sent message. Otherwise, it moves one step ahead and repeats the second step of the decoding

¹⁴In the literature this decoder is often referred to as the “strong typicality” decoder.

procedure on the basis of y_{t-n+2}^{t+1} , i.e., it tests whether y_{t-n+2}^{t+1} is typical with a codeword.

At the end of the asynchronism time window, i.e., at time $A_n + n - 1$, if $\hat{P}_{A_n+n-1}^{A_n+n-1}$ is either a noisy type or if it is typical with none of the codewords, the decoder declares a message at random.

Throughout the argument we assume that the typicality parameter μ is a negligible, strictly positive quantity.

We first show that, on average, a randomly chosen codebook combined with the sequential decoding procedure described above achieves the rate-exponent pairs (R, α) claimed by the theorem. This, as we show at the end of the proof, implies the existence of a nonrandom codebook that, together with the above decoding procedure, achieves any pair (R, α) claimed by the theorem.

Let $\ln M/n = I(PQ) - \epsilon$, $\epsilon > 0$. We first compute the average, over messages and codes, expected reaction delay and probability of error. These quantities, by symmetry of the encoding and decoding procedures, are the same as the average over codes expected reaction delay and probability of error conditioned on the sending of a particular message. Below, expected reaction delay and error probability are computed conditioned on the sending of message $m = 1$.

Define the following events:

$$\begin{aligned}\mathcal{E}_1 &= \{D(\hat{P}_{Y_{\nu+n-1}}^{\nu+n-1} \| Q_*) < \alpha, \text{ i.e., } \hat{P}_{Y_{\nu+n-1}}^{\nu+n-1} \text{ is a "noise type"}\}, \\ \mathcal{E}_2 &= \{Y_{\nu+n-1}^{\nu+n-1} \text{ is not typical with } C^n(1)\}, \\ \mathcal{E}_3 &= \{D(\hat{P}_{Y_{t-n+1}}^t \| Q_*) \geq \alpha \text{ for some } t < \nu\}.\end{aligned}$$

For the reaction delay we have

$$\begin{aligned}\mathbb{E}_1(\tau_n - \nu)^+ &= \mathbb{E}_1[(\tau_n - \nu)^+ \mathbb{1}(\tau_n \geq \nu + 2n)] \\ &\quad + \mathbb{E}_1[(\tau_n - \nu)^+ \mathbb{1}(\nu + n \leq \tau_n < \nu + 2n)] \\ &\quad + \mathbb{E}_1[(\tau_n - \nu)^+ \mathbb{1}(\tau_n < \nu + n)] \\ &\leq (A_n + n - 1)\mathbb{P}_1(\tau_n \geq \nu + 2n) \\ &\quad + 2n\mathbb{P}_1(\nu + n \leq \tau_n < \nu + 2n) + n,\end{aligned}\tag{19}$$

where the subscript 1 in \mathbb{E}_1 and \mathbb{P}_1 indicates conditioning on the event that message $m = 1$ is sent. The two probability terms on the right-hand side of the second inequality of (19) are bounded as follows.

The term $\mathbb{P}_1(\tau_n \geq \nu + 2n)$ is upper bounded by the probability that the decoding window starts after time $\nu + n - 1$. This, in turn, is upper bounded by the probability of the event that, at time $\nu + n - 1$, the last n output symbols induce a noise type. Therefore, we have

$$\begin{aligned}\mathbb{P}_1(\tau_n \geq \nu + 2n) &\leq \mathbb{P}_1(\mathcal{E}_1) \\ &\leq \sum_{\{V \in \mathcal{P}_n^y: D(V \| Q_*) \leq \alpha\}} e^{-nD(V \| (PQ)_y)} \\ &\leq \sum_{\{V \in \mathcal{P}_n^y: D(V \| Q_*) \leq \alpha\}} e^{-n\alpha} \\ &\leq \text{poly}(n)e^{-n\alpha},\end{aligned}\tag{20}$$

where the second inequality follows from the definition of the event \mathcal{E}_1 and Fact 2; where the third inequality follows from

(17) (which implies that if $D(V \| Q_*) \leq \alpha$ then necessarily $D(V \| (PQ)_y) \geq \alpha$); and where the fourth inequality follows from Fact 1.

The probability $\mathbb{P}_1(\nu + n \leq \tau_n < \nu + 2n)$ is at most the probability that the decoder has not stopped by time $\nu + n - 1$. This probability, in turn, is at most the probability that, at time $\nu + n - 1$, the last n output symbols either induce a noisy type, or are not typical with the sent codeword $C^n(1)$ (recall that message $m = 1$ is sent). By union bound we get

$$\begin{aligned}\mathbb{P}_1(\nu + n \leq \tau_n < \nu + 2n) &\leq \mathbb{P}_1(\tau_n \geq \nu + n) \\ &\leq \mathbb{P}_1(\mathcal{E}_1) + \mathbb{P}_1(\mathcal{E}_2) \\ &\leq \text{poly}(n)e^{-n\alpha} + o(1) \\ &= o(1) \quad (n \rightarrow \infty),\end{aligned}\tag{21}$$

where we used the last three computation steps of (20) to bound $\mathbb{P}_1(\mathcal{E}_1)$, and where we used [7, Lemma 2.12, p. 34] to show that $\mathbb{P}_1(\mathcal{E}_2)$ tends to zero as n tends to infinity. From (19), (20), and (21), we deduce that

$$\mathbb{E}_1(\tau_n - \nu)^+ \leq n(1 + o(1)) \quad (n \rightarrow \infty)$$

since $A_n = e^{n(\alpha - \epsilon)}$, by assumption.

We now compute $\mathbb{P}_1(\mathcal{E})$, the average error probability conditioned on sending message $m = 1$. We have

$$\begin{aligned}\mathbb{P}_1(\mathcal{E}) &= \mathbb{P}_1(\mathcal{E} \cap \{\tau_n < \nu\}) \\ &\quad + \mathbb{P}_1(\mathcal{E} \cap \{\nu \leq \tau_n \leq \nu + n - 1\}) \\ &\quad + \mathbb{P}_1(\mathcal{E} \cap \{\tau_n \geq \nu + n\}) \\ &\leq \mathbb{P}_1(\tau_n < \nu) + \mathbb{P}_1(\mathcal{E} \cap \{\nu \leq \tau_n \leq \nu + n - 1\}) \\ &\quad + \mathbb{P}_1(\tau_n \geq \nu + n) \\ &\leq \mathbb{P}_1(\mathcal{E}_3) + \mathbb{P}_1(\mathcal{E} \cap \{\nu \leq \tau_n \leq \nu + n - 1\}) \\ &\quad + o(1) \quad (n \rightarrow \infty),\end{aligned}\tag{22}$$

where for the last inequality we used the definition of \mathcal{E}_3 and upper bounded $\mathbb{P}_1(\tau \geq \nu + n)$ using the last three computation steps of (21).

For $\mathbb{P}_1(\mathcal{E}_3)$, we have

$$\begin{aligned}\mathbb{P}_1(\mathcal{E}_3) &= \mathbb{P}(\cup_{t < \nu} \{D(\hat{P}_{Y_{t-n+1}}^t \| Q_*) \geq \alpha\}) \\ &\leq A_n \sum_{\{V \in \mathcal{P}_n^x: D(V \| Q_*) \geq \alpha\}} e^{-nD(V \| Q_*)} \\ &\leq A_n \sum_{\{V \in \mathcal{P}_n^x: D(V \| Q_*) \geq \alpha\}} e^{-n\alpha} \\ &\leq A_n e^{-n\alpha} \text{poly}(n) \\ &= o(1) \quad (n \rightarrow \infty)\end{aligned}\tag{23}$$

where the first inequality in (23) follows from the union bound over time and Fact 2; where the third inequality follows from Fact 1; and where the last equality holds since $A_n = e^{n(\alpha - \epsilon)}$, by assumption.

We now show that

$$\mathbb{P}_1(\mathcal{E} \cap \{\nu \leq \tau_n \leq \nu + n - 1\}) = o(1) \quad (n \rightarrow \infty),\tag{24}$$

which, together with (22) and (23), shows that $\mathbb{P}_1(\mathcal{E})$ goes to zero as $n \rightarrow \infty$.

We have

$$\begin{aligned}
& \mathbb{P}_1(\mathcal{E} \cap \{\nu \leq \tau_n \leq \nu + n - 1\}) \\
&= \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-1} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3\}) \\
&\quad + \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-1} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}) \\
&\leq \mathbb{P}_1(\mathcal{E}_3) + \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-1} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}) \\
&\leq o(1) + \mathbb{P}_1(\{\mathcal{E} \cap \{\tau_n = \nu + n - 1\}\}) \\
&\quad + \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-2} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}) \\
&\leq o(1) + o(1) \\
&\quad + \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-2} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}) \quad (n \rightarrow \infty)
\end{aligned} \tag{25}$$

where the second inequality follows from (23); where the fourth inequality follows from the definition of event \mathcal{E}_2 ; and where the third inequality follows from the fact that, given the correct codeword location, i.e., $\tau_n = \nu + n - 1$, the typicality decoder guarantees vanishing error probability since we assumed that $\ln M/n = I(PQ) - \epsilon$ (see [7, Chapter 2.1]).

The event $\{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}$, with $\nu \leq t \leq \nu + n - 2$, happens when a block of n consecutive symbols, received between $\nu - n + 1$ and $\nu + n - 2$, is jointly typical with a codeword other than the sent codeword $C^n(1)$. Consider a block Y^n in this range, and let $J \in \mathcal{P}_n^{\mathcal{X}, \mathcal{Y}}$ be a typical joint type, i.e.

$$|J(x, y) - P(x)Q(y|x)| \leq \mu$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ —recall that $\mu > 0$ is the typicality parameter, which we assume to be a negligible quantity throughout the proof.

For some $1 \leq k \leq n - 1$, the first k symbols of block Y^n are generated by noise, and the remaining $n - k$ symbols are generated by the sent codeword, i.e., corresponding to $m = 1$. Thus, Y^n is independent of any unsent codeword $C^n(m)$. The probability that $C^n(m)$, $m \neq 1$, together with Y^n yields a particular type J is upper bounded as follows:

$$\begin{aligned}
& \mathbb{P}(\hat{P}_{C^n(m), Y^n} = J) \\
&= \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}(Y^n = y^n) \sum_{x^n: \hat{P}_{x^n, y^n} = J} \mathbb{P}(X^n = x^n) \\
&= \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}(Y^n = y^n) \sum_{x^n: \hat{P}_{x^n, y^n} = J} e^{-n(H(J_X) + D(J_X \| P))} \\
&\leq \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}(Y^n = y^n) e^{-nH(J_X)} |\{x^n: \hat{P}_{x^n, y^n} = J\}| \\
&\leq \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}_1(Y^n = y^n) e^{-nH(J_X)} e^{nH(J_{X|Y})} \\
&\leq e^{-nI(J)},
\end{aligned} \tag{26}$$

where $H(J_X)$ denotes the entropy of the left marginal of J ,

$$H(J_{X|Y}) \triangleq - \sum_{y \in \mathcal{Y}} J_Y(y) \sum_{x \in \mathcal{X}} J_{X|Y}(x|y) \ln J_{X|Y}(x|y),$$

and where $I(J)$ denotes the mutual information induced by J .

The first equality in (26) follows from the independence of $C^n(m)$ and Y^n , the second equality follows from [11,

Theorem 11.1.2, p. 349], and the second inequality follows from [7, Lemma 2.5, p. 31].

It follows that the probability that an unsent codeword $C^n(m)$ together with Y^n yields a type J that is typical, i.e., close to PQ , is upper bounded as

$$\mathbb{P}_1(\hat{P}_{C^n(m), Y^n} = J) \leq e^{-n(I(PQ) - \epsilon/2)}$$

for all n large enough, by continuity of the mutual information.¹⁵

Note that the set of inequalities (26) holds for any block of n consecutive output symbols Y^n that is independent of codeword $C^n(m)$.¹⁶ Hence, from the union bound, it follows that

$$\begin{aligned}
& \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-2} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}) \\
&\leq n \sum_{m \neq 1} \sum_{\{J \in \mathcal{P}_{\mathcal{X}, \mathcal{Y}}^n: \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \\
&\quad |J(x, y) - P(x)Q(y|x)| \leq \mu\}} \mathbb{P}(\hat{P}_{C^n(m), Y^n} = J) \\
&\leq n M e^{-n(I(PQ) - \epsilon/2)} \text{poly}(n) \\
&\leq e^{-n\epsilon/2} \text{poly}(n),
\end{aligned} \tag{27}$$

where the second inequality follows from Fact 1, and where the third inequality follows from the assumption that $\ln M/n = I(PQ) - \epsilon$. Combining (27) with (25) yields (24).

So far, we have proved that a random codebook has a decoding delay averaged over messages that is at most $n(1 + o(1))$ ($n \rightarrow \infty$), and an error probability averaged over messages that vanishes as $n \rightarrow \infty$, whenever $A_n = e^{n(\alpha - \epsilon)}$, $\epsilon > 0$. This, as we now show, implies the existence of nonrandom codebooks achieving the same performance, yielding the desired result. The expurgation arguments we use are standard and in the same spirit as those given in [11, p. 203-204] or [12, p. 151].

For a particular codebook \mathcal{C}_n , let $\mathbb{P}(\mathcal{E}|\mathcal{C}_n)$ and $\mathbb{E}((\tau_n - \nu)^+|\mathcal{C}_n)$ be the average, over messages, error probability and reaction delay, respectively. We have proved that for any $\epsilon > 0$,

$$\mathbb{E}(\mathbb{E}(\tau_n - \nu)^+|\mathcal{C}_n)) \leq n(1 + \epsilon)$$

and

$$\mathbb{E}(\mathbb{P}(\mathcal{E}|\mathcal{C}_n)) \leq \epsilon$$

for all n large enough.

Define events

$$\mathcal{A}_1 = \{\mathbb{E}(\tau_n - \nu)^+|\mathcal{C}_n) \leq n(1 + \epsilon)^2\},$$

and

$$\mathcal{A}_2 = \{\mathbb{P}(\mathcal{E}|\mathcal{C}_n) \leq \epsilon k\}$$

where k is arbitrary.

From Markov's inequality it follows that¹⁷

$$\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2) \geq 1 - \frac{1}{1 + \epsilon} - \frac{1}{k}.$$

¹⁵The typicality parameter $\mu = \mu(\epsilon) > 0$ is chosen small enough so that this inequality holds.

¹⁶Note that the fact that Y^n is partly generated by noise and partly by the sent codeword $C^n(1)$ is not used to establish (26).

¹⁷Probability here is averaged over randomly generated codewords.

Letting k be large enough so that the right-hand side of the above inequality is positive, we deduce that there exists a particular code \mathcal{C}_n such that

$$\mathbb{E}(\tau_n - \nu)^+ | \mathcal{C}_n \leq n(1 + \epsilon)^2$$

and

$$\mathbb{P}(\mathcal{E} | \mathcal{C}_n) \leq \epsilon k.$$

We now remove from \mathcal{C}_n codewords with poor reaction delay and error probability. Repeating the argument above with the fixed code \mathcal{C}_n , we see that a positive fraction of the codewords of \mathcal{C}_n have expected decoding delay at most $n(1 + \epsilon)^3$ and error probability at most ϵk^2 . By only keeping this set of codewords, we conclude that for any $\epsilon > 0$ and all n large enough, there exists a rate $R = I(PQ) - \epsilon$ code operating at asynchronism level $A = e^{(\alpha - \epsilon)n}$ with maximum error probability less than ϵ . ■

Remark 2: It is possible to somewhat strengthen the conclusion of Theorem 2 in two ways. First, it can be strengthened by observing that what we actually proved is that the error probability not only vanishes but does so exponentially in n .¹⁸ Second, it can be strengthened by showing that the proposed random coding scheme achieves (6) with equality. A proof is deferred to Appendix A.

B. Proof of Theorem 3

We show that any rate $R > 0$ coding scheme operates at an asynchronism α bounded from above by $\max_{\mathcal{S}} \min\{\alpha_1, \alpha_2\}$, where \mathcal{S} , α_1 , and α_2 are defined in the theorem's statement.

We prove Theorem 3 by establishing the following four claims.

The first claim says that, without loss of generality, we may restrict ourselves to constant composition codes. Specifically, it is possible to expurgate an arbitrary code to make it of constant composition while impacting (asymptotically) neither the rate nor the asynchronism exponent the original code is operating at. In more detail, the expurgated codebook is such that all codewords have the same type, and also so that all codewords have the same type over the first Δ_n symbols (recall that $\Delta_n \triangleq \max_m \mathbb{E}(\tau_n - \nu)^+$). The parameter δ in Theorem 3 corresponds to the ratio Δ_n/n , and P_1 and P_2 correspond to the empirical types over the first Δ_n symbols and the whole codeword (all n symbols), respectively.

Fix an arbitrarily small constant $\epsilon > 0$.

Claim 1: Given any coding scheme $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ achieving (R, α) with $R > 0$ and $\alpha > 0$, there exists a second coding scheme $\{(\mathcal{C}'_n, (\tau_n, \phi_n))\}_{n \geq 1}$ achieving (R, α) that is obtained by expurgation, i.e., $\mathcal{C}'_n \subset \mathcal{C}_n$, $n = 1, 2, \dots$, and that has constant composition with respect to some distribution P_n^1 over the first

$$d(n) \triangleq \min\{[(1 + \epsilon)\Delta_n], n\} \quad (28)$$

symbols, and constant composition with respect to some distribution P_n^2 over n symbols. (Hence, if $[(1 + \epsilon)\Delta_n] \geq n$,

¹⁸Note that the error probability of the typicality decoder given the correct message location, i.e., $\mathbb{P}(\mathcal{E} \cap \{\tau_n = \nu + n - 1\})$, is exponentially small in n [7, Chapter 2].

then $P_n^1 = P_n^2$.) Distributions P_n^1 and P_n^2 satisfy Claims 2–4 below.

Distribution P_n^1 plays the same role as the codeword distribution for synchronous communication. As such it should induce a large enough input-output channel mutual information to support rate R communication.

Claim 2: For all n large enough

$$R \leq I(P_n^1 Q)(1 + \epsilon).$$

Distribution P_n^2 is specific to asynchronous communication. Intuitively, P_n^2 should induce an output distribution that is sufficiently different from pure noise so that to allow a decoder to distinguish between noise and any particular transmitted message when the asynchronism level corresponds to α . Proper message detection means that the decoder should not overreact to a sent codeword (i.e., declare a message before even it is sent), but also not miss the sent codeword. As an extreme case, it is possible to achieve a reaction delay $\mathbb{E}(\tau - \nu)^+$ equal to zero by setting $\tau = 1$, at the expense of a large probability of error. In contrast, one clearly minimizes the error probability by waiting until the end of the asynchronism window, i.e., by setting $\tau = A_n + n - 1$, at the expense of the rate, which will be negligible in this case.

The ability to properly detect only a single codeword with type P_n^2 is captured by condition $\alpha \leq \alpha_2$ where α_2 is defined in the theorem's statement. This condition is equivalently stated as:

Claim 3: For any $W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ and for all n large enough, at least one of the following two inequalities holds

$$\begin{aligned} \alpha &< D(W \| Q_\star | P_n^2) + \epsilon, \\ \alpha &< D(W \| Q | P_n^2) + \epsilon. \end{aligned}$$

As it turns out, if the synchronization threshold is finite, P_n^1 plays also a role in the decoder's ability to properly detect the transmitted message. This is captured by condition $\alpha \leq \alpha_1$ where α_1 is defined in the theorem's statement. Intuitively, α_1 relates to the probability that the noise produces a string of length n that looks typical with the output of a *randomly selected codeword*. If $\alpha > \alpha_1$, the noise produces many such strings with high probability, which implies a large probability of error.

Claim 4: For all n large enough,

$$\alpha \leq \frac{d(n)}{n} (I(P_n^1 Q) - R + D((P_n^1 Q)_Y \| Q_\star)) + \epsilon$$

provided that $\alpha_o < \infty$.

Note that, by contrast with the condition in Claim 3, the condition in Claim 4 depends also on the communication rate since the error yielding to the latter condition depends on the number of codewords.

Before proving the above claims, we show how they imply Theorem 3. The first part of the Theorem, i.e., when $\alpha_o < \infty$, follows from Claims 1-4. To see this, note that the bounds α_1 and α_2 in the Theorem correspond to the bounds of Claims 3 and 4, respectively, maximized over P_n^1 and P_n^2 . The maximization is subjected to the two constraints given by Claims 1 and 2: P_n^1 and P_n^2 are the empirical distributions of the codewords of \mathcal{C}'_n over the first δn symbols ($\delta \in [0, 1]$), and

over the entire codeword length, respectively, and condition $R \leq I(P_n^1 Q)(1 + \epsilon)$ must be satisfied. Since $\epsilon > 0$ is arbitrary, the result then follows by taking the limit $\epsilon \downarrow 0$ on the above derived bound on α .

Similarly, the second part of Theorem 3, i.e., when $\alpha_o = \infty$, is a consequence of Claim 3 only.

We now prove the claims. As above, $\epsilon > 0$ is supposed to be an arbitrarily small constant.

Proofs of Claims 1 and 2: We show that for all n large enough, we have

$$\frac{R - \epsilon}{1 + \epsilon} \leq \frac{\ln |\mathcal{C}'_n|}{d(n)} \leq I(P_n^1 Q) + \epsilon, \quad (29)$$

where \mathcal{C}'_n is a subset of codewords from \mathcal{C}_n that have constant composition P_n^1 over the first $d(n)$ symbols, where $d(n)$ is defined in (28), and constant composition P_n^2 over n symbols. This is done via an expurgation argument in the spirit of [12, p. 151] and [11, p. 203-204].

We first show the left-hand side inequality of (29). Since $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ achieves a rate R , by definition (see Definition 1) we have

$$\frac{\ln |\mathcal{C}_n|}{\Delta_n} \geq R - \epsilon/2$$

for all n large enough. Therefore,

$$\frac{\ln |\mathcal{C}_n|}{d(n)} \geq \frac{R - \epsilon/2}{1 + \epsilon}$$

for all n large enough.

Now, group the codewords of \mathcal{C}_n into families such that elements of the same family have the same type over the first $d(n)$ symbols. Let \mathcal{C}''_n be the largest such family and let P_n^1 be its type. Within \mathcal{C}''_n , consider the largest subfamily \mathcal{C}'_n of codewords that have constant composition over n symbols, and let P_n^2 be its type (hence, all the codewords in \mathcal{C}'_n have common type P_n^1 over $d(n)$ symbols and common type P_n^2 over n symbols).

By assumption, $R > 0$, so \mathcal{C}_n has a number of codewords that is exponential in Δ_n . Due to Fact 1, to establish the left-hand side inequality of (29), i.e., to show that \mathcal{C}'_n achieves essentially the same rate as \mathcal{C}_n , it suffices to show that the number of subfamilies in \mathcal{C}'_n is bounded by a polynomial in Δ_n . We do this assuming that $\alpha_o < \infty$ and that Claim 4 (to be proved) holds.

By assumption, $\alpha_o < \infty$, and thus from Theorem 1 we have that $D((PQ)_y \| Q_*) < \infty$ for any input distribution P . Using Claim 4 and the assumption that $\alpha > 0$, we deduce that $\liminf_{n \rightarrow \infty} d(n)/n > 0$, which implies that n cannot grow faster than linearly in Δ_n . Therefore, Fact 1 implies that the number of subfamilies of \mathcal{C}'_n is bounded by a polynomial in Δ_n .

We now prove the right-hand side inequality of (29). Letting \mathcal{E}^c denote the event of a correct decoding, Markov's inequality

implies that for every message index m ,

$$\begin{aligned} \mathbb{P}_m(\{(\tau_n - \nu)^+ \leq (1 + \epsilon)\Delta_n\} \cap \mathcal{E}^c) \\ \geq 1 - \frac{\mathbb{E}_m(\tau_n - \nu)^+}{\Delta_n} \frac{1}{1 + \epsilon} - \mathbb{P}_m(\mathcal{E}) \\ \geq 1 - \frac{1}{1 + \epsilon} - \mathbb{P}_m(\mathcal{E}), \end{aligned} \quad (30)$$

since $\Delta_n \triangleq \max_m \mathbb{E}_m(\tau_n - \nu)^+$. The right-hand side of (30) is strictly greater than zero for n large enough because an (R, α) coding scheme achieves a vanishing maximum error probability as $n \rightarrow \infty$. This means that \mathcal{C}'_n is a good code for the synchronous channel, i.e., for $A = 1$. More precisely, the codebook formed by truncating each codeword in \mathcal{C}'_n to include only the first $d(n)$ symbols achieves a probability of error (asymptotically) bounded away from one with a suitable decoding function. This implies that the right-hand side of (29) holds for n large enough by [7, Corollary 1.4, p. 104]. ■

In establishing the remaining claims of the proof, unless otherwise stated, whenever we refer to a codeword it is assumed to belong to codebook \mathcal{C}'_n . Moreover, for convenience, and with only minor abuse of notation, we let M denote the number of codewords in \mathcal{C}'_n .

Proof of Claim 3: We fix $W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ and show that for all n large enough, at least one of the two inequalities

$$D(W \| Q | P_n^2) > \alpha - \epsilon,$$

$$D(W \| Q_* | P_n^2) > \alpha - \epsilon,$$

must hold. To establish this, it may be helpful to interpret W as the true channel behavior during the information transmission period, i.e., as the conditional distribution induced by the transmitted codeword and the corresponding channel output. With this interpretation, $D(W \| Q | P_n^2)$ represents the large deviation exponent of the probability that the underlying channel Q behaves as W when codeword distribution is P_n^2 , and $D(W \| Q_* | P_n^2)$ represents the large deviation exponent of the probability that the noise behaves as W when codeword distribution is P_n^2 . As it turns out, if both the above inequalities are reversed for a certain W , the asynchronism exponent is too large. In fact, in this case both the transmitted message and pure noise are very likely to produce such a W . This, in turn will confuse the decoder. It will either miss the transmitted codeword or stop before even the actual codeword is sent.

In the sequel, we often use the shorthand notation $\mathcal{J}_W(m)$ for $\mathcal{J}_W^n(c^n(m))$.

Observe first that if n is such that

$$\mathbb{P}_m(Y_\nu^{\nu+n-1} \in \mathcal{J}_W(m)) = 0, \quad (31)$$

then

$$D(W \| Q | P_n^2) = \infty,$$

by Fact 3. Similarly, observe that if n is such that

$$\mathbb{P}_*(Y_\nu^{\nu+n-1} \in \mathcal{J}_W(m)) = 0, \quad (32)$$

where \mathbb{P}_* denotes the probability under pure noise (i.e., the Y_i 's are i.i.d. according to Q_*), then

$$D(W \| Q_* | P_n^2) = \infty.$$

Since the above two observations hold regardless of m (because all codewords in \mathcal{C}'_n have the same type), Claim 3 holds trivially for any value of n for which (31) or (32) is satisfied.

In the sequel, we thus restrict our attention to values of n for which

$$\mathbb{P}_m(Y_\nu^{\nu+n-1} \in \mathcal{T}_W(m)) \neq 0 \quad (33)$$

and

$$\mathbb{P}_*(Y_\nu^{\nu+n-1} \in \mathcal{T}_W(m)) \neq 0. \quad (34)$$

Our approach is to use a change of measure to show that if Claim 3 does not hold, then the expected reaction delay grows exponentially with n , implying that the rate is asymptotically equal to zero. To see this, note that any coding scheme that achieves vanishing error probability cannot have $\ln M$ grow faster than linearly with n , simply because of the limitations imposed by the capacity of the synchronous channel. Therefore, if $\mathbb{E}(\tau_n - \nu)^+$ grows exponentially with n , the rate goes to zero exponentially with n . And note that for $\mathbb{E}(\tau_n - \nu)^+$ to grow exponentially, it suffices that $\mathbb{E}_m(\tau_n - \nu)^+$ grows exponentially for at least one message index m , since $\Delta_n = \max_m \mathbb{E}_m(\tau_n - \nu)^+$ by definition.

To simplify the exposition and avoid heavy notation, in the following arguments we disregard discrepancies due to the rounding of noninteger quantities. We may, for instance, treat A/n as an integer even if A is not a multiple of n . This has no consequences on the final results, as these discrepancies vanish when we consider code with blocklength n tending to infinity.

We start by lower bounding the reaction delay as¹⁹

$$\begin{aligned} \Delta_n &\triangleq \max_m \frac{1}{A} \sum_{t=1}^A \mathbb{E}_{m,t}(\tau_n - t)^+ \\ &\geq \frac{1}{3} \sum_{t=1}^{A/3} \mathbb{P}_{m,t}((\tau_n - t)^+ \geq A/3) \\ &\geq \frac{1}{3} \sum_{t=1}^{A/3} \mathbb{P}_{m,t}(\tau_n \geq t + A/3) \\ &\geq \frac{1}{3} \sum_{t=1}^{A/3} \mathbb{P}_{m,t}(\tau_n \geq 2A/3), \end{aligned} \quad (35)$$

where for the first inequality we used Markov's inequality. The message index m on the right-hand side of (35) will be specified later; for now it may correspond to any message.

We lower bound each term $\mathbb{P}_{m,t}(\tau_n \geq 2A/3)$ in the above sum as

$$\begin{aligned} \mathbb{P}_{m,t}(\tau_n \geq 2A/3) &\geq \mathbb{P}_{m,t}(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ &\quad \times \mathbb{P}_{m,t}(Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ &\geq \mathbb{P}_{m,t}(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ &\quad \times e^{-nD_1} \text{poly}(n), \end{aligned} \quad (36)$$

¹⁹Recall that the subscripts m, t indicate conditioning on the event that message m starts being sent at time t .

where $D_1 \triangleq D(W \| Q | P_n^2)$, and where the second inequality follows from Fact 3.²⁰

The key step is to apply the change of measure

$$\begin{aligned} \mathbb{P}_{m,t}(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ = \mathbb{P}_*(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)). \end{aligned} \quad (37)$$

To see that (37) holds, first note that for any y^n

$$\begin{aligned} \mathbb{P}_{m,t}(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} = y^n) \\ = \mathbb{P}_*(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} = y^n) \end{aligned}$$

since distribution $\mathbb{P}_{m,t}$ and \mathbb{P}_* differ only over channel outputs Y_t^{t+n-1} .

Next, since sequences inside $\mathcal{T}_W(m)$ are permutations of each other

$$\begin{aligned} \mathbb{P}_{m,t}(Y_t^{t+n-1} = y^n \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) &= \frac{1}{|\mathcal{T}_W(m)|} \\ &= \mathbb{P}_*(Y_t^{t+n-1} = y^n \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)), \end{aligned}$$

we get

$$\begin{aligned} \mathbb{P}_{m,t}(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ = \sum_{y^n \in \mathcal{T}_W(m)} \mathbb{P}_{m,t}(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} = y^n) \\ \quad \times \mathbb{P}_{m,t}(Y_t^{t+n-1} = y^n \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ = \sum_{y^n \in \mathcal{T}_W(m)} \mathbb{P}_*(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} = y^n) \\ \quad \times \mathbb{P}_*(Y_t^{t+n-1} = y^n \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ = \mathbb{P}_*(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)). \end{aligned}$$

This proves (37). Substituting (37) into the right-hand side of (36) and using (35), we get

$$\begin{aligned} \Delta_n &\geq e^{-nD_1} \text{poly}(n) \\ &\quad \times \sum_{t=1}^{A/3} \mathbb{P}_*(\tau_n \geq 2A/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ &\geq e^{-n(D_1+D_2)} \text{poly}(n) \\ &\quad \times \sum_{t=1}^{A/3} \mathbb{P}_*(\tau_n \geq 2A/3, Y_t^{t+n-1} \in \mathcal{T}_W(m)), \end{aligned}$$

where $D_2 \triangleq D(W \| Q_* | P_n^2)$, and where the last inequality follows from Fact 3. By summing only over the indices that are multiples of n , we obtain the weaker inequality

$$\begin{aligned} \Delta_n &\geq e^{-n(D_1+D_2)} \text{poly}(n) \\ &\quad \times \sum_{j=1}^{A/3n} \mathbb{P}_*(\tau_n \geq 2A/3, Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)). \end{aligned} \quad (38)$$

Using (38), we show that $\mathbb{E}(\tau_n - \nu)^+$ grows exponentially with n whenever D_1 and D_2 are both upper bounded by $\alpha - \epsilon$. This, as we saw above, implies that the rate is asymptotically equal to zero, yielding Claim 3.

²⁰Note that the right-hand side of the first inequality in (36) is well-defined because of (33).

Let $A = e^{\alpha n}$, and let $\mu \triangleq \epsilon/2$. We rewrite the above summation over $A/3n$ indices as a sum of $A_1 = e^{n(\alpha - D_2 - \mu)}/3n$ superblocks of $A_2 = e^{n(D_2 + \mu)}$ indices. We have

$$\begin{aligned} & \sum_{j=1}^{A/3n} \mathbb{P}_*(\tau_n \geq 2A_n/3, Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)) \\ &= \sum_{s=1}^{A_1} \sum_{j \in I_s} \mathbb{P}_*(\tau_n \geq 2A_n/3, Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)), \end{aligned}$$

where I_s denotes the s th superblock of A_2 indices. Applying the union bound (in reverse), we see that

$$\begin{aligned} & \sum_{s=1}^{A_1} \sum_{j \in I_s} \mathbb{P}_*(\tau_n \geq 2A_n/3, Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)) \\ & \geq \sum_{s=1}^{A_1} \mathbb{P}_*\left(\tau_n \geq 2A_n/3, \cup_{j \in I_s} \{Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)\}\right). \end{aligned}$$

We now show that each term

$$\mathbb{P}_*(\tau_n \geq 2A_n/3, \cup_{j \in I_s} \{Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)\}) \quad (39)$$

in the above summation is large, say greater than $1/2$, by showing that each of them involves the intersection of two large probability events. This, together with (38), implies that

$$\begin{aligned} \Delta_n &= \text{poly}(n) \Omega(e^{n(\alpha - D_1 - \mu)}) \\ &\geq \Omega(\exp(n\epsilon/2)) \end{aligned} \quad (40)$$

since $D_1 \leq \alpha - \epsilon$, yielding the desired result.²¹

Letting \mathcal{E} denote the decoding error event, we have for all n large enough

$$\begin{aligned} \epsilon &\geq \mathbb{P}_m(\mathcal{E}) \\ &\geq \mathbb{P}_m(\mathcal{E} | \nu > 2A_n/3, \tau_n \leq 2A_n/3) \\ &\quad \times \mathbb{P}_m(\nu > 2A_n/3, \tau_n \leq 2A_n/3) \\ &\geq \frac{1}{2} \mathbb{P}_m(\nu > 2A_n/3) \mathbb{P}_m(\tau_n \leq 2A_n/3 | \nu > 2A_n/3) \\ &\geq \frac{1}{6} \mathbb{P}_m(\tau_n \leq 2A_n/3 | \nu > 2A_n/3). \end{aligned} \quad (41)$$

The third inequality follows by noting that the event $\{\nu > 2A_n/3, \tau_n \leq 2A_n/3\}$ corresponds to the situation where the decoder stops after observing only pure noise. Since a codebook consists of at least two codewords,²² such an event causes an error with probability at least $1/2$ for at least one message m . Thus, inequality (41) holds under the assumption that m corresponds to such a message.²³

²¹Our proof shows that for all indices n for which $D_1 \leq \alpha - \epsilon$ and $D_2 \leq \alpha - \epsilon$, (40) holds. Therefore, if $D_1 \leq \alpha - \epsilon$ and $D_2 \leq \alpha - \epsilon$ for every n large enough, the reaction delay grows exponentially with n , and thus the rate vanishes. In the case where $D_1 \leq \alpha - \epsilon$ and $D_2 \leq \alpha - \epsilon$ does not hold for all n large enough, but still holds for infinitely many values of n , the corresponding asymptotic rate is still zero by Definition 1.

²²By assumption, see Section III.

²³Regarding the fourth inequality in (41), note that $\mathbb{P}_m(\nu > 2A_n/3)$ should be lower bounded by $1/4$ instead of $1/3$ had we taken into account discrepancies due to rounding of noninteger quantities. As mentioned earlier, we disregard these discrepancies as they play no role asymptotically.

Since the event $\{\tau_n \leq 2A_n/3\}$ depends on the channel outputs only up to time $2A_n/3$, we have

$$\mathbb{P}_m(\tau_n \leq 2A_n/3 | \nu > 2A_n/3) = \mathbb{P}_*(\tau_n \leq 2A_n/3). \quad (42)$$

Combining (42) with (41) we get

$$\mathbb{P}_*(\tau_n > 2A_n/3) \geq 1 - 6\epsilon. \quad (43)$$

Now, because the Y_{jn}^{jn+n-1} , $j \in I_s$, are i.i.d. under \mathbb{P}_* ,

$$\begin{aligned} & \mathbb{P}_*\left(\cup_{j \in I_s} \{Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)\}\right) \\ &= 1 - (1 - \mathbb{P}_*(Y^n \in \mathcal{T}_W(m)))^{|I_s|}. \end{aligned}$$

From Fact 3 it follows that

$$\mathbb{P}_*(Y^n \in \mathcal{T}_W(m)) \geq \text{poly}(n) \exp(-nD_2),$$

and by definition $|I_s| = e^{n(D_2 + \mu)}$, so

$$\mathbb{P}_*\left(\cup_{j \in I_s} \{Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)\}\right) = 1 - o(1) \quad (n \rightarrow \infty). \quad (44)$$

Combining (43) and (44), we see that each term (39) involves the intersection of large probability events for at least one message index m . For such a message index, by choosing ϵ sufficiently small, we see that for all sufficiently large n , every single term (39), $s \in \{1, 2, \dots, A_1\}$ is bigger than $1/2$. ■

Finally, to establish the remaining Claim 4, we make use of Theorem 5, whose proof is provided in Appendix B. This theorem implies that any nontrivial codebook contains a (large) set of codewords whose rate is almost the same as the original codebook and whose error probability decays faster than polynomially, say as $e^{-\sqrt{n}}$, with a suitable decoder. Note that we don't use the full implication of Theorem 5.

Proof of Claim 4: The main idea behind the proof is that if Claim 4 does not hold, the noise is likely to produce an output that is “typical” with a codeword before the message is even sent, which means that any decoder must have large error probability. Although the idea is fairly simple, it turns out that a suitable definition for “typical” set and its related error probability analysis make the proof somewhat lengthy.

Proceeding formally, consider inequality (30). This inequality says that, with nonzero probability, the decoder makes a correct decision and stops soon after the beginning of the information transmission period. This motivates the definition of a new random process, which we call the modified output process. With a slight abuse of notation, in the remainder of the proof we use $Y_1, Y_2, \dots, Y_{A+n-1}$ to denote the modified output process. The modified output process is generated as if the sent codeword were truncated at the position $\nu + d(n)$, where $d(n)$ is defined in (28). Hence, this process can be thought of as the random process “viewed” by the sequential decoder.

Specifically, the distribution of the modified output process is as follows. If

$$n \geq \lfloor \Delta_n(1 + \epsilon) \rfloor,$$

then the Y_i 's for

$$i \in \{1, \dots, \nu - 1\} \cup \{\nu + \lfloor \Delta_n(1 + \epsilon) \rfloor, \dots, A_n + n - 1\}$$

are i.i.d. according to Q_* , whereas the block

$$Y_\nu, Y_{\nu+1}, \dots, Y_{\nu+\lfloor \Delta_n(1+\epsilon) \rfloor - 1}$$

is distributed according to $Q(\cdot|c^{d(n)})$, the output distribution given that a *randomly selected* codeword has been transmitted. Note that, in the conditioning, we use $c^{d(n)}$ instead of $c^{d(n)}(m)$ to emphasize that the output distribution is averaged over all possible messages, i.e., by definition

$$Q(y^{d(n)}|c^{d(n)}) = \frac{1}{M} \sum_{m=1}^M Q(y^{d(n)}|c^{d(n)}(m)).$$

Instead, if

$$n < \lfloor \Delta_n(1+\epsilon) \rfloor,$$

then the modified output process has the same distribution as the original one, i.e., the Y_i 's for

$$i \in \{1, \dots, \nu - 1\} \cup \{\nu + n, \dots, A_n + n - 1\}$$

are i.i.d. according to Q_* , whereas the block

$$Y_\nu, Y_{\nu+1}, \dots, Y_{\nu+n-1}$$

is distributed according to $Q(\cdot|c^n)$.

Consider the following augmented decoder that, in addition to declaring a message, also outputs the time interval

$$[\tau_n - \lfloor \Delta_n(1+\epsilon) \rfloor + 1, \tau_n - \lfloor \Delta_n(1+\epsilon) \rfloor + 2, \dots, \tau_n],$$

of size $\lfloor \Delta_n(1+\epsilon) \rfloor$. A simple consequence of the right-hand side of (30) being (asymptotically) bounded away from zero is that, for n large enough, if the augmented decoder is given a modified output process instead of the original one, with a strictly positive probability it declares the correct message, and the time interval it outputs contains ν .

Now, suppose the decoder is given the modified output process and that it is revealed that the (possibly truncated) sent codeword was sent in one of the

$$r_n = \left\lfloor \frac{(A_n + n - 1) - (\nu \bmod d(n))}{d(n)} \right\rfloor \quad (45)$$

consecutive blocks of duration $d(n)$, as shown in Fig. 12. Using this additional knowledge, the decoder can now both declare the sent message and output a list of

$$\ell_n = \lceil \lfloor \Delta_n(1+\epsilon) \rfloor / d(n) \rceil \quad (46)$$

block positions, one of which corresponding to the sent message, with a probability strictly away from zero for all n large enough. To do this the decoder, at time τ_n , declares the decoded message and declares the ℓ_n blocks that overlap with the time indices in

$$\{\tau_n - \lfloor \Delta_n(1+\epsilon) \rfloor + 1, \tau_n - \lfloor \Delta_n(1+\epsilon) \rfloor + 2, \dots, \tau_n\}.$$

We now show that the above task that consists of declaring the sent message and producing a list of ℓ_n blocks of size $d(n)$, one of which being the output of the transmitted message, can be performed only if α satisfies Claim 4. To that aim we consider the performance of the (optimal) maximum likelihood decoder that observes output sequences of maximal length

$$d(n) \cdot r_n.$$



Fig. 12. Parsing of the entire received sequence of size $A + n - 1$ into r_n blocks of length $d(n)$, one of which being generated by the sent message, and the others being generated by noise.

Given a sample $y_1, y_2, \dots, y_{A+n-1}$ of the modified output process, and its parsing into consecutive blocks of duration $d(n)$, the optimal decoder outputs a list of ℓ_n blocks that are most likely to occur. More precisely, the maximum likelihood ℓ_n -list decoder operates as follows. For each message m , it finds a list of ℓ_n blocks $y^{d(n)}$ (among all r_n blocks) that maximize the ratio

$$\frac{Q(y^{d(n)}|c^{d(n)}(m))}{Q(y^{d(n)}|\star)}, \quad (47)$$

and computes the sum of these ratios. The maximum likelihood ℓ_n -list decoder then outputs the list whose sum is maximal, and declares the corresponding message.²⁴

The rest of the proof consists in deriving an upper bound on the probability of correct maximum likelihood ℓ_n -list decoding, and show that this bound tends to zero if Claim 4 is not satisfied. To that aim, we first quantify the probability that the noise distribution Q_* outputs a sequence that is typical with a codeword, since the performance of the maximum likelihood ℓ_n -list decoder depends on this probability, as we show below.

By assumption, $(\mathcal{C}'_n, (\tau_n, \phi_n))$ achieves a probability of error $\epsilon'_n \rightarrow 0$ as $n \rightarrow \infty$ at the asynchronism exponent α . This implies that \mathcal{C}'_n can also achieve a nontrivial error probability on the synchronous channel (i.e., with $A = 1$). Specifically, by using the same argument as for (30), we deduce that we can use \mathcal{C}'_n on the synchronous channel, force decoding to happen at the fixed time

$$d(n) = \min\{n, \lfloor (1+\epsilon)\Delta_n \rfloor\},$$

where Δ_n corresponds to the reaction delay obtained by $(\mathcal{C}'_n, (\tau_n, \phi_n))$ in the asynchronous setting, and guarantee a (maximum) probability of error ϵ''_n such that

$$\epsilon''_n \leq \frac{1}{1+\epsilon} + \epsilon'_n$$

with a suitable decoder. Since the right-hand side of the above inequality is strictly below one for n large enough, Theorem 5 with $q = 1/4$ implies that the code \mathcal{C}'_n has a large subcode $\tilde{\mathcal{C}}_n$, i.e., of almost the same rate with respect to $d(n)$, that, together with an appropriate decoding function $\tilde{\phi}_n$, achieves a maximum error probability at most equal to

$$\epsilon_n = 2(n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|} \exp(-\sqrt{n}/(2 \ln 2)) \quad (48)$$

for all n large enough.

²⁴To see this, consider a channel output $y^{d(n) \cdot r_n}$ that is composed of r_n consecutive blocks of size $d(n)$, where the j th block is generated by codeword $c^{d(n)}$ and where all the other blocks are generated by noise. The probability of this channel output is

$$\mathbb{P}(y^{d(n) \cdot r_n} | m, j) = Q(y^{d(n)}(j) | c^{d(n)}) \prod_{i \neq j} Q_*(y^{d(n)}(i))$$

where $y^{d(n)}(j)$, $j \in \{1, 2, \dots, r_n\}$, denotes the j th bloc of $y^{d(n) \cdot r_n}$

We now start a digression on the code $(\tilde{\mathcal{C}}_n, \tilde{\phi}_n)$ when used on channel Q synchronously. The point is to exhibit a set of “typical output sequences” that cause the decoder $\tilde{\phi}_n$ to make an error with “large probability.” We then move back to the asynchronous channel Q and show that when Claim 4 does not hold, the noise distribution Q_* is likely to produce typical output sequences, thereby inducing the maximum likelihood ℓ_n -list decoder into error.

Unless stated otherwise, we now consider $(\tilde{\mathcal{C}}_n, \tilde{\phi}_n)$ when used on the synchronous channel. In particular error events are defined with respect to this setting.

The set of typical output sequences is obtained through a few steps. We first define the set \mathcal{A}_m with respect to codeword $c^{d(n)}(m) \in \tilde{\mathcal{C}}_n$ as

$$\mathcal{A}_m = \{y^{d(n)} \in \mathcal{T}_W(c^{d(n)}(m)) \text{ with } W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}} : \mathbb{P}(\mathcal{T}_W(c^{d(n)}(m)) | c^{d(n)}(m)) \geq \sqrt{\epsilon_{d(n)}}\} \quad (49)$$

where ϵ_n is defined in (48).

Note that, by using Fact 3, it can easily be checked that \mathcal{A}_m is nonempty for n large enough (depending on $|\mathcal{X}|$ and $|\mathcal{Y}|$), which we assume throughout the argument. For a fixed m , consider the set of sequences in \mathcal{A}_m that maximize (47). These sequences form a set $\mathcal{T}_{\bar{Q}}(c^{d(n)}(m))$, for some $\bar{Q} \in \mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}$. It follows that for every message index m for which $c^{d(n)}(m) \in \tilde{\mathcal{C}}_n$, we have

$$\begin{aligned} \epsilon_{d(n)} &\geq \mathbb{P}_m(\mathcal{E}) \\ &\geq \mathbb{P}_m(\mathcal{E} | \{Y_\nu^{\nu+d(n)-1} \in \mathcal{T}_{\bar{Q}}(c^{d(n)}(m))\}) \\ &\times \mathbb{P}_m(\{Y_\nu^{\nu+d(n)-1} \in \mathcal{T}_{\bar{Q}}(c^{d(n)}(m))\}) \\ &\geq \mathbb{P}_m(\mathcal{E} | \{Y_\nu^{\nu+d(n)-1} \in \mathcal{T}_{\bar{Q}}(c^{d(n)}(m))\}) \sqrt{\epsilon_{d(n)}} \\ &\geq \mathbb{P}_m(\mathcal{E} | \{Y_\nu^{\nu+d(n)-1} \in \mathcal{B}_m\}) \times \\ &\mathbb{P}_m(\{Y_\nu^{\nu+d(n)-1} \in \mathcal{B}_m\} | \{Y_\nu^{\nu+d(n)-1} \in \mathcal{T}_{\bar{Q}}(c^{d(n)}(m))\}) \\ &\times \sqrt{\epsilon_{d(n)}} \\ &\geq \frac{1}{2} \times \\ &\mathbb{P}_m(\{Y_\nu^{\nu+d(n)-1} \in \mathcal{B}_m\} | \{Y_\nu^{\nu+d(n)-1} \in \mathcal{T}_{\bar{Q}}(c^{d(n)}(m))\}) \\ &\times \sqrt{\epsilon_{d(n)}} \end{aligned} \quad (50)$$

where for the third inequality we used the definition of \bar{Q} ; where on the right-hand side of the fourth inequality we defined the set

$$\mathcal{B}_m \triangleq \{y^{d(n)} \in \mathcal{T}_{\bar{Q}}(c^{d(n)}(m)) \cap (\cup_{m' \neq m} \mathcal{T}_{\bar{Q}}(c^{d(n)}(m')))\};$$

and where the fifth inequality follows from this definition.²⁵

From (50) we get

$$\begin{aligned} \mathbb{P}_m(\{Y_\nu^{\nu+d(n)-1} \in \mathcal{B}_m\} | \{Y_\nu^{\nu+d(n)-1} \in \mathcal{T}_{\bar{Q}}(c^{d(n)}(m))\}) \\ \leq 2\sqrt{\epsilon_{d(n)}}. \end{aligned} \quad (51)$$

²⁵Note that, given that message m is sent, if the channel produces a sequence in \mathcal{B}_m at its output, the (standard) optimal maximum likelihood decoder makes an error with probability at least half. Hence the decoding rule $\tilde{\phi}_n$ also makes an error with probability at least half.

Therefore, by defining $\tilde{\mathcal{B}}_m$ as

$$\tilde{\mathcal{B}}_m \triangleq \mathcal{T}_{\bar{Q}}(c^{d(n)}(m)) \setminus \mathcal{B}_m$$

the complement of \mathcal{B}_m in $\mathcal{T}_{\bar{Q}}(c^{d(n)}(m))$, it follows from (51) that

$$|\tilde{\mathcal{B}}_m| > (1 - 2\sqrt{\epsilon_{d(n)}})|\mathcal{T}_{\bar{Q}}(c^{d(n)}(m))|,$$

since under \mathbb{P}_m all the sequences in $\mathcal{T}_{\bar{Q}}(c^{d(n)}(m))$ are equiprobable.

The set $\cup_{m' \neq m}^M \tilde{\mathcal{B}}_{m'}$ is the sought set of “typical output sequences” that causes the decoder make an error with “high probability” conditioned on the sending of message m and conditioned on the channel outputting a sequence in $\mathcal{T}_{\bar{Q}}(c^{d(n)}(m))$. This ends our digression on $(\tilde{\mathcal{C}}_n, \tilde{\phi}_n)$.

We now compute a lower bound on the probability under Q_* of producing a sequence in $\cup_{m=2}^M \tilde{\mathcal{B}}_m$. Because the sets $\{\tilde{\mathcal{B}}_m\}$ are disjoint, we deduce that

$$\begin{aligned} |\cup_{m=2}^M \tilde{\mathcal{B}}_m| &\geq (1 - 2\sqrt{\epsilon_n}) \sum_{m=2}^M |\mathcal{T}_{\bar{Q}}(c^{d(n)}(m))| \\ &\geq \frac{(1 - 2\sqrt{\epsilon_n})}{(d(n) + 1)^{|\mathcal{X}| \cdot |\mathcal{Y}|}} (M - 1) e^{d(n)H(\bar{Q}|P_n^1)} \\ &\geq \frac{1}{(4n)^{|\mathcal{X}| \cdot |\mathcal{Y}|}} e^{d(n)(H(\bar{Q}|P_n^1) + \ln M/d(n))} \end{aligned} \quad (52)$$

for all n large enough. For the second inequality we used [7, Lemma 2.5, p. 31]. For the third inequality we used the fact that $d(n) \leq n$, $M \geq 2$, $(1 - 2\sqrt{\epsilon_{d(n)}}) \geq 1/2$ for n large enough,²⁶ and that, without loss of generality, we may assume that $|\mathcal{X}| \cdot |\mathcal{Y}| \geq 2$ since the synchronous capacity C is non-zero—as we assume throughout the paper. Hence we get

$$\begin{aligned} Q_*(\cup_{m=2}^M \tilde{\mathcal{B}}_m) &= \sum_{y^{d(n)} \in \cup_{m=2}^M \tilde{\mathcal{B}}_m} Q_*(y^{d(n)}) \\ &\geq |\cup_{m=2}^M \tilde{\mathcal{B}}_m| \min_{y^{d(n)} \in \cup_{m=2}^M \tilde{\mathcal{B}}_m} Q_*(y^{d(n)}) \\ &\geq \frac{1}{(4n)^{|\mathcal{X}| \cdot |\mathcal{Y}|}} e^{d(n)(H(\bar{Q}|P_n^1) + (\ln M)/d(n))} \\ &\quad \times e^{-d(n)(D((P_n^1 \bar{Q})_{\mathcal{Y}} \| Q_*) + H((P_n^1 \bar{Q})_{\mathcal{Y}}))} \end{aligned}$$

for all n large enough, where for the second inequality we used (52) and [11, Theorem 11.1.2, p. 349]. Letting

$$e_n \triangleq \ln I(P_n^1 \bar{Q}) - (\ln M)/d(n) + D((P_n^1 \bar{Q})_{\mathcal{Y}} \| Q_*),$$

we thus have

$$Q_*(\cup_{m=2}^M \tilde{\mathcal{B}}_m) \geq \frac{1}{(4n)^{|\mathcal{X}| \cdot |\mathcal{Y}|}} e^{-e_n \cdot d(n)} \quad (53)$$

for n large enough.

Using (53), we now prove Claim 4 by contradiction. Specifically, assuming that

$$\alpha > \frac{d(n)}{n} e_n + \epsilon/2 \quad \text{for infinitely many indices } n, \quad (54)$$

we prove that, given message $m = 1$ is sent, the probability of error of the maximum likelihood ℓ_n -list decoder does not

²⁶Note that $d(n) \xrightarrow{n \rightarrow \infty} \infty$ since the coding scheme under consideration achieves a strictly positive rate.

converge to zero. As final step, we prove that the opposite of (54) implies Claim 4.

Define the events

$$\begin{aligned}\mathcal{E}_1 &= \{Y_\nu^{\nu+n-1} \notin \mathcal{A}_1\}, \\ \mathcal{E}_2 &= \{Z \leq \frac{1}{2} \frac{1}{(4n)^{2|\mathcal{X}||\mathcal{Y}|}} e^{\alpha n - e_n \cdot d(n)}\},\end{aligned}$$

where \mathcal{A}_1 is defined in (49), and where Z denotes the random variable that counts the number of blocks generated by Q_\star that are in $\cup_{m=2}^M \mathcal{B}_m$. Define also the complement set

$$\mathcal{E}_3 \triangleq (\mathcal{E}_1 \cup \mathcal{E}_2)^c.$$

The probability that the maximum likelihood ℓ_n -list decoder makes a *correct* decision given that message $m = 1$ is sent is upper bounded as

$$\begin{aligned}\mathbb{P}_1(\mathcal{E}^c) &= \sum_{i=1}^3 \mathbb{P}_1(\mathcal{E}^c | \mathcal{E}_i) \mathbb{P}_1(\mathcal{E}_i) \\ &\leq \mathbb{P}_1(\mathcal{E}_1) + \mathbb{P}_1(\mathcal{E}_2) + \mathbb{P}_1(\mathcal{E}^c | \mathcal{E}_3).\end{aligned}\quad (55)$$

From the definition of \mathcal{A}_1 , we have

$$\mathbb{P}_1(\mathcal{E}_1) = o(1) \quad (n \rightarrow \infty). \quad (56)$$

Now for $\mathbb{P}_1(\mathcal{E}_2)$. There are $r_n - 1$ blocks independently generated by Q_\star (r_n is defined in (45)). Each of these blocks has a probability at least equal to the right-hand side of (53) to fall within $\cup_{m=2}^M \mathcal{B}_m$. Hence, using (53) we get

$$\begin{aligned}\mathbb{E}_1 Z &\geq (r_n - 1) \frac{1}{(4n)^{|\mathcal{X}||\mathcal{Y}|}} e^{-e_n d(n)} \\ &\geq \frac{1}{(4n)^{2|\mathcal{X}||\mathcal{Y}|}} e^{\alpha n - e_n d(n)}\end{aligned}\quad (57)$$

since $r_n \geq e^{\alpha n}/n$. Therefore,

$$\begin{aligned}\mathbb{P}_1(\mathcal{E}_2) &\leq \mathbb{P}_1(Z \leq (\mathbb{E}_1 Z)/2) \\ &\leq \frac{4}{\mathbb{E}_1 Z} \\ &\leq \text{poly}(n) e^{-\alpha n + e_n d(n)}\end{aligned}\quad (58)$$

where the first inequality follows from (57) and the definition of \mathcal{E}_2 ; where for the second inequality we used Chebyshev's inequality and the fact that the variance of a binomial is upper bounded by its mean; and where for the third inequality we used (57).

Finally for $\mathbb{P}_1(\mathcal{E}^c | \mathcal{E}_3)$. Given \mathcal{E}_3 , the decoder sees at least

$$\frac{1}{2} \frac{1}{(4n)^{2|\mathcal{X}||\mathcal{Y}|}} e^{\alpha n - e_n \cdot d(n)}$$

time slots whose corresponding ratios (47) are at least as large as the one induced by the correct block $Y_\nu^{\nu+d(n)-1}$. Hence, given \mathcal{E}_3 , the decoder produces a list of ℓ_n block positions, one of which corresponds to the sent message, with probability at most

$$\begin{aligned}\mathbb{P}_1(\mathcal{E}^c | \mathcal{E}_3) &\leq \ell_n \left(\frac{1}{2} \frac{1}{(4n)^{2|\mathcal{X}||\mathcal{Y}|}} e^{\alpha n - e_n \cdot d(n)} \right)^{-1} \\ &= \text{poly}(n) e^{-\alpha n + e_n \cdot d(n)},\end{aligned}\quad (59)$$

where the first inequality follows from union bound, and where for the equality we used the fact that finite rate implies $\ell_n = \text{poly}(n)$.²⁷

From (55), (56), (58), and (59), the probability that the maximum likelihood ℓ_n -list decoder makes a correct decision, $\mathbb{P}_1(\mathcal{E}^c)$, is arbitrarily small for infinitely many indices n whenever (54) holds. Therefore to achieve vanishing error probability we must have, for all n large enough,

$$\begin{aligned}\alpha &\leq \\ &\frac{d(n)}{n} (I(P_n^1 \bar{Q}) - (\ln M)/d(n) + D((P_n^1 \bar{Q})_{\mathcal{Y}} \| Q_\star)) \\ &\quad + \epsilon/2.\end{aligned}\quad (60)$$

We now show, via a continuity argument, that the above condition implies Claim 4. Recall that $\bar{Q} \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, defined just after (49), depends on n and has the property

$$\mathbb{P}(\mathcal{T}_{\bar{Q}}(c^{d(n)}(m) | c^{d(n)}(m))) \geq \sqrt{\epsilon_{d(n)}}. \quad (61)$$

Now, from Fact 3 we also have the upper bound

$$\mathbb{P}(\mathcal{T}_{\bar{Q}}(c^{d(n)}(m) | c^{d(n)}(m))) \leq e^{-d(n)D(\bar{Q} \| Q | P_n^1)}. \quad (62)$$

Since $\sqrt{\epsilon_{d(n)}} = \Omega(e^{-\sqrt{d(n)}})$, from (61) and (62) we get

$$D(\bar{Q} \| Q | P_n^1) \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and therefore

$$\|P_n^1 \bar{Q} - P_n^1 Q\| \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

where $\|\cdot\|$ denotes the L_1 norm. Hence, by continuity of the divergence, condition (60) gives, for all n large enough,

$$\alpha \leq \quad (63)$$

$$\begin{aligned}&\frac{d(n)}{n} (I(P_n^1 Q) - (\ln M)/d(n) + D((P_n^1 Q)_{\mathcal{Y}} \| Q_\star)) \\ &\quad + \epsilon\end{aligned}\quad (64)$$

which yields Claim 4. \blacksquare

C. Proof of Corollary 3

By assumption α_o is nonzero since divergence is always non-negative. This implies that the synchronous capacity is nonzero by the last claim of Theorem 1. This, in turn, implies that (R, α) is achievable for some sufficiently small $R > 0$ and $\alpha > 0$ by [3, Corollary 1].

Using Theorem 3,

$$\alpha \leq \alpha(R) \leq \max_s \alpha_2 \quad (65)$$

where α_2 is given by expression (10). In this expression, by letting $W = Q_\star$ in the minimization, we deduce that $\alpha_2 \leq D(Q_\star \| Q | P_2)$, and therefore

$$\begin{aligned}\max_s \alpha_2 &\leq \max_s D(Q_\star \| Q | P_2) \\ &= \max_{P_2} D(Q_\star \| Q | P_2) \\ &= \max_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} Q_\star(y) \ln \frac{Q_\star(y)}{Q(y|x)} \\ &= \max_{x \in \mathcal{X}} D(Q_\star \| Q(\cdot|x)),\end{aligned}$$

²⁷This follows from the definition of rate $R = \ln M / \mathbb{E}(\tau - \nu)^+$, the fact that $\ln M/n \leq C$ for reliable communication, and the definition of ℓ_n (46).

and from (65) we get

$$\alpha \leq \max_{x \in \mathcal{X}} D(Q_\star \| Q(\cdot|x)).$$

Since, by assumption,

$$\alpha_o > \max_{x \in \mathcal{X}} D(Q_\star \| Q(\cdot|x)),$$

and since $\alpha_o = \alpha(R=0)$ by Theorem 1, it follows that $\alpha(R)$ is discontinuous at $R=0$. ■

D. Proof of Theorem 4

We first exhibit a coding scheme that achieves any (R, α) with $R \leq C$ and

$$\alpha \leq \max_{P \in \mathcal{P}^{\mathcal{X}}} \min_{W \in \mathcal{P}^{\mathcal{Y}|X}} \max\{D(W \| Q|P), D(W \| Q_\star|P)\}.$$

All codewords start with a common preamble that is composed of $(\ln(n))^2$ repetitions of a symbol x such that $D(Q(\cdot|x) \| Q_\star) = \infty$ (such a symbol exists since $\alpha_o = \infty$). The next $(\ln(n))^3$ symbols of each codeword are drawn from a code that achieves a rate equal to $R - \epsilon$ on the synchronous channel. Finally, all the codewords end with a common large suffix s^l of size $l = n - (\ln(n))^2 - (\ln(n))^3$ that has an empirical type P such that, for all $W \in \mathcal{P}^{\mathcal{Y}|X}$, at least one of the following two inequalities holds:

$$\begin{aligned} D(W \| Q|P) &\geq \alpha \\ D(W \| Q_\star|P) &\geq \alpha. \end{aligned}$$

The receiver runs two sequential decoders in parallel, and makes a decision whenever one of the two decoder declares a message. If the two decoders declare different messages at the same time, the receiver declares one of the messages at random.

The first decoder tries to identify the sent message by first locating the preamble. At time t it checks if the channel output y_t can be generated by x but cannot be generated by noise, i.e., if

$$Q(y_t|x) > 0 \quad \text{and} \quad Q(y_t|\star) = 0. \quad (66)$$

If condition (66) does not hold, the decoder moves one-step ahead and checks condition (66) at time $t+1$. If condition (66) does hold, the decoder marks the current time as the beginning of the “decoding window” and proceeds to the second step. The second step consists in exactly locating and identifying the sent codeword. Once the beginning of the decoding window has been marked, the decoder makes a decision the first time it observes $(\ln n)^3$ symbols that are typical with one of the codewords. If no such time is found within $(\ln(n))^2 + (\ln(n))^3$ time steps from the time the decoding window has been marked, the decoder declares a random message.

The purpose of the second decoder is to control the average reaction delay by stopping the decoding process in the rare event when the first decoder misses the codeword. Specifically, the second “decoder” is only a stopping rule based on the suffix s^l . At each time t the second decoder checks whether $D(\hat{P}_{Y_{t-l+1}^t} \| Q|P) < \alpha$. If so, the decoder stops and declares a random message. If not, the decoder moves one step ahead.

The arguments for proving that the coding scheme described above achieves (R, α) provided

$$\alpha \leq \max_P \min_W \max\{D(W \| Q|P), D(W \| Q_\star|P)\} \quad (67)$$

closely parallel those used to prove Theorem 2, and are therefore omitted.²⁸

The converse is the second part of Theorem 3. ■

E. Proof of Theorem 6

1) *Lower bound:* To establish the lower bound in Theorem 6, we exhibit a training based scheme with preamble size ηn with

$$\eta = (1 - R/C), \quad (68)$$

and that achieves any rate asynchronism pair (R, α) such that

$$\alpha \leq m_1 \left(1 - \frac{R}{C}\right) \quad R \in (0, C] \quad (69)$$

where

$$m_1 \triangleq \max_{P \in \mathcal{P}^{\mathcal{X}}} \min_{W \in \mathcal{P}^{\mathcal{Y}|X}} \max\{D(W \| Q|P), D(W \| Q_\star|P)\}.$$

Fix $R \in (0, C]$ and let α satisfy (69). Each codeword starts with a common preamble of size ηn where η is given by (68) and whose empirical distribution is equal to²⁹

$$\begin{aligned} P_p &\triangleq \\ \arg \max_{P \in \mathcal{P}^{\mathcal{X}}} &\left(\min_{W \in \mathcal{P}^{\mathcal{Y}|X}} \max\{D(W \| Q|P), D(W \| Q_\star|P)\} \right). \end{aligned}$$

The remaining $(1 - \eta)n$ symbols of each codeword are i.i.d. generated according to a distribution P that almost achieves capacity of the synchronous channel, i.e., such that $I(PQ) = C - \epsilon$ for some small $\epsilon > 0$.

Note that by (69) and (68), α is such that for any $W \in \mathcal{P}^{\mathcal{Y}|X}$ at least one of the following two inequalities holds:

$$\begin{aligned} D(W \| Q|P_p) &\geq \alpha/\eta \\ D(W \| Q_\star|P_p) &\geq \alpha/\eta. \end{aligned} \quad (70)$$

The preamble detection rule is to stop the first time when last ηn output symbols $Y_{t-\eta n+1}^t$ induce an empirical conditional probability $\hat{P}_{Y_{t-\eta n+1}^t|x^{\eta n}}$ such that

$$D(\hat{P}_{Y_{t-\eta n+1}^t|x^{\eta n}} \| Q|P_p) \leq D(\hat{P}_{Y_{t-\eta n+1}^t|x^{\eta n}} \| Q_\star|P_p) \quad (71)$$

where $x^{\eta n}$ is the preamble.

When the preamble is located, the decoder makes a decision on the basis of the upcoming $(1 - \eta)n$ output symbols using maximum likelihood decoding. If no preamble has been located by time $A_n + n - 1$, the decoder declares a message at random.

We compute the reaction delay and the error probability. For notational convenience, instead of the decoding time, we consider the time τ_n that the decoder detects the preamble, i.e., the first time t such that (71) holds. The actual decoding

²⁸In particular, note that the first decoder never stops before time ν .

²⁹ P_p need not be a valid type for finite values of n , but this small discrepancy plays no role asymptotically since P_p can be approximated arbitrarily well with types of order sufficiently large.

time occurs $(1-\eta)n$ time instants after the preamble has been detected, i.e., at time $\tau_n + (1-\eta)n$.

For the reaction delay we have

$$\begin{aligned}\mathbb{E}(\tau_n - \nu)^+ &= \mathbb{E}_1(\tau_n - \nu)^+ \\ &= \mathbb{E}_1[(\tau_n - \nu)^+ \mathbb{1}(\tau_n \geq \nu + \eta n)] \\ &\quad + \mathbb{E}_1[(\tau_n - \nu)^+ \mathbb{1}(\tau_n \leq \nu + \eta n - 1)] \\ &\leq (A_n + n - 1)\mathbb{P}_1(\tau_n \geq \nu + \eta n) + \eta n\end{aligned}\quad (72)$$

where, as usual, the subscript 1 in \mathbb{E}_1 and \mathbb{P}_1 indicates conditioning on the event that message $m = 1$ is sent. A similar computation as in (20) yields

$$\begin{aligned}\mathbb{P}_1(\tau_n \geq \nu + \eta n) &\leq \mathbb{P}_1(D(\hat{P}_{Y^{\nu+\eta n-1}|x^{\eta n}}||Q|P_p) \geq \alpha/\eta) \\ &\leq \sum_{W \in \mathcal{P}_n^{\mathcal{Y}|X}: D(W||Q|P_p) \geq \alpha/\eta} e^{-\eta n D(W||Q|P_p)} \\ &\leq \text{poly}(n)e^{-n\alpha}.\end{aligned}\quad (73)$$

The first inequality follows from the fact that event $\{\tau_n \geq \nu + n\}$ is included into event

$$\{D(\hat{P}_{Y^{\nu+\eta n-1}|x^{\eta n}}||Q|P_p) > D(\hat{P}_{Y^{\nu+\eta n-1}|x^{\eta n}}||Q_\star|P_p)\}$$

which, in turn, is included into event

$$\{D(\hat{P}_{Y^{\nu+\eta n-1}|x^{\eta n}}||Q|P_p) \geq \alpha/\eta\}$$

because of (70). The second inequality follows from Fact 2. Hence, from (72) and (73)

$$\mathbb{E}(\tau_n - \nu)^+ \leq \eta n + o(1) \quad (74)$$

whenever $A_n = e^{n(\alpha-\epsilon)}$, $\epsilon > 0$. Since the actual decoding time occurs $(1-\eta)n$ time instants after τ_n , where $\eta = (1-R/C)$, and that the code used to transmit information achieves the capacity of the synchronous channel, the above strategy operates at rate R .

To show that the above strategy achieves vanishing error probability, one uses arguments similar to those used to prove Theorem 2 (see from paragraph after (21) onwards), so the proof is omitted. There is one little caveat in the analysis that concerns the event when the preamble is located somewhat earlier than its actual timing, i.e., when the decoder locates the preamble over a time period $[t - \eta n + 1, \dots, t]$ with $\nu \leq t \leq \nu + \eta n - 2$. One way to make the probability of this event vanish as $n \rightarrow \infty$, is to have the preamble have a ‘‘sufficiently large’’ Hamming distance with any of its shifts. To guarantee this, one just needs to modify the original preamble in a few (say, $\log n$) positions. This modifies the preamble type negligibly. For a detailed discussion on how to make this modification, we refer the reader to [9], where the problem is discussed in the context of sequential frame synchronization.

Each instance of the above random coding strategy satisfies the conditions of Definition 3; there is a common preamble of size ηn and the decoder decides to stop at any particular time t based on $Y_{t-n+1}^{t-n+\eta n}$. We now show that there exists a particular instance yielding the desired rate and error probability.

First note that the above rate analysis only depends on the preamble, and not on the codebook that follows the preamble.

Hence, because the error probability, averaged over codebooks and messages, vanishes, we deduce that there exists at least one codebook that achieves rate R and whose average over messages error probability tends to zero.

From this code, we remove codewords with poor error probability, say whose error probabilities are at least twice the average error probability. The resulting expurgated code has a rate that tends to R and a vanishing maximum error probability.

2) *Upper bound*: To establish the upper bound it suffices to show that for training based schemes (R, α) with $R > 0$ must satisfy

$$\alpha \leq m_2 \left(1 - \frac{R}{C}\right). \quad (75)$$

The upper bound in Theorem 6 then follows from (75) and the general upper bound derived in Theorem 3.

The upper bound (75) follows from the following lemma:

Lemma 1: A rate $R > 0$ coding scheme whose decoder operates according to a sliding window stopping rule with window size ηn cannot achieve an asynchronism exponent larger than ηm_2 .

Lemma 1 says that any coding scheme with a limited memory stopping rule capable of processing only ηn symbols at a time achieves an asynchronism exponent at most $O(\eta)$, unless $R = 0$ or if the channel is degenerate, i.e., $\alpha_o = m_2 = \infty$, in which case Lemma 1 is trivial and we have the asynchronous capacity expression given by Theorem 4.

To deduce (75) from Lemma 1, consider a training-based scheme which achieves a delay Δ with a non-trivial error probability (i.e., bounded away from 0). Because the preamble conveys no information, the rate is at most

$$C \frac{\min\{\Delta, n\} - \eta n}{\Delta} \leq C(1 - \eta)$$

by the channel coding theorem for a synchronous channel. Hence, for a rate $R > 0$ training-based scheme the training fraction η is upper bounded as

$$\eta \leq 1 - \frac{R}{C}.$$

This implies (75) by Lemma 1. ■

Proof of Lemma 1: The lemma holds trivially if $m_2 = \infty$. We thus assume that $m_2 < \infty$. Consider a training-based scheme $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ in the sense of Definition 3. For notational convenience, we consider τ_n to be the time when the decoder detects the preamble. The actual decoding time (in the sense of Definition 3 part 2) occurs $(1-\eta)n$ times instants after the preamble has been detected, i.e., at time $\tau_n + (1-\eta)n$. This allows us to write τ_n as

$$\tau_n = \inf\{t \geq 1 : S_t = 1\},$$

where

$$S_t = S_t(Y_{t-\eta n+1}^t) \quad 1 \leq t \leq A_n + n - 1,$$

referred to as the ‘‘stopping rule at time t ,’’ is a binary random variable such that $\{S_t = 1\}$ represents the set of output sequences $y_{t-\eta n+1}^t$ which make τ_n stop at time t , assuming that τ_n hasn’t stopped before time t .

Now, every sequence $y^{\eta n} \in \mathcal{Y}^{\eta n}$ satisfies

$$Q_*(y^{\eta n}) \geq e^{-m_2 \eta n}.$$

Therefore, any deterministic stopping rule stops at any particular time either with probability zero or with probability at least $e^{-m_2 \eta n}$, i.e., for all t , either the stopping rule S_t satisfies $\mathbb{P}(S_t = 1) \geq e^{-m_2 \eta n}$ or it is trivial in the sense that $\mathbb{P}(S_t = 1) = 0$. For now, we assume that the stopping rule is deterministic; the randomized case follows easily as we describe at the end of the proof.

Let \mathcal{S} denote the subset of indices $t \in \{1, 2, \dots, A_n/4\}$ such that S_t is non-trivial, and let $\bar{\mathcal{S}}_k$ denote the subset of indices in \mathcal{S} that are congruent to $k \bmod \eta n$, i.e.,

$$\bar{\mathcal{S}}_k = \{t : t \in \mathcal{S}, t = j \cdot \eta n + k, j = 0, 1, \dots\}.$$

Note that for each k , the set of stopping rules S_t , $t \in \bar{\mathcal{S}}_k$ are independent since S_t depends only on $Y_{t-\eta n+1}^t$.

By repeating the same argument as in (41)-(42), for any $\epsilon > 0$, for all n large enough and any message index m the error probability $\mathbb{P}_m(\mathcal{E})$ satisfies

$$\begin{aligned} \epsilon &\geq \mathbb{P}_m(\mathcal{E}) \\ &\geq \frac{1}{4} \mathbb{P}_*(\tau_n \leq A_n/2). \end{aligned} \quad (76)$$

Since $\epsilon > 0$ is arbitrary, we deduce

$$\mathbb{P}_*(\tau_n \geq A_n/2) \geq 1/2 \quad (77)$$

i.e., a coding scheme achieves a vanishing error probability only if the probability of stopping after time $A_n/2$ is at least 0.5 when the channel input is all \star 's. Thus, assuming that our coding scheme achieves vanishing error probability, we have

$$|\mathcal{S}| < \eta n e^{m_2 \eta n}.$$

To see this, note that if $|\mathcal{S}| \geq \eta n e^{m_2 \eta n}$, then there exists a value k^* such that $|\bar{\mathcal{S}}_{k^*}| \geq e^{m_2 \eta n}$, and hence

$$\begin{aligned} \mathbb{P}_*(\tau_n \geq A_n/2) &\leq \mathbb{P}_*(S_t = 0, t \in \mathcal{S}) \\ &\leq \mathbb{P}_*(S_t = 0, t \in \bar{\mathcal{S}}_{k^*}) \\ &= (1 - e^{-m_2 \eta n})^{|\bar{\mathcal{S}}_{k^*}|} \\ &\leq (1 - e^{-m_2 \eta n})^{e^{m_2 \eta n}}. \end{aligned}$$

Since the above last term tends to $1/e < 1/2$ for n large enough, $\mathbb{P}_*(\tau_n \geq A_n/2) < 1/2$ for n large enough, which is in conflict with the assumption that the coding scheme achieves vanishing error probability.

The fact that $|\mathcal{S}| < \eta n e^{m_2 \eta n}$ implies, as we shall prove later, that

$$\mathbb{P}(\tau_n \geq A_n/2 | \nu \leq A_n/4) \geq \frac{1}{2} \left(1 - \frac{8\eta^2 n^2 e^{m_2 \eta n}}{A_n} \right). \quad (78)$$

Hence,

$$\begin{aligned} \mathbb{E}(\tau_n - \nu)^+ &\geq \mathbb{E}((\tau_n - \nu)^+ | \tau_n \geq A_n/2, \nu \leq A_n/4) \\ &\quad \times \mathbb{P}(\tau_n \geq A_n/2, \nu \leq A_n/4) \\ &\geq \frac{A_n}{16} \mathbb{P}(\tau_n \geq A_n/2 | \nu \leq A_n/4) \\ &\geq \frac{A_n}{32} \left(1 - \frac{8\eta^2 n^2 e^{m_2 \eta n}}{A_n} \right). \end{aligned} \quad (79)$$

where for the second inequality we used the fact that ν is uniformly distributed, and where the third inequality holds by (78). Letting $A_n = e^{\alpha n}$, from (79) we deduce that if $\alpha > m\eta$, then $\mathbb{E}(\tau_n - \nu)^+$ grows exponentially with n , implying that the rate is asymptotically zero.³⁰ Hence a sliding window stopping rule which operates on a window of size ηn cannot accommodate a positive rate while achieving an asynchronism exponent larger than ηm . This establishes the desired result.

We now show (78). Let \mathcal{N} be the subset of indices in $\{1, 2, \dots, A_n/4\}$ with the following property. For any $t \in \mathcal{N}$, the $2n$ indices $\{t, t+1, \dots, t+2n-1\}$ do not belong to \mathcal{S} , i.e., all $2n$ of the associated stopping rules are trivial. Then we have

$$\begin{aligned} \mathbb{P}(\tau_n \geq A_n/2 | \nu \leq A_n/4) &\geq \mathbb{P}(\tau_n \geq A_n/2 | \nu \in \mathcal{N}) \\ &\quad \times \mathbb{P}(\nu \in \mathcal{N} | \nu \leq A_n/4) \\ &= \mathbb{P}(\tau_n \geq A_n/2 | \nu \in \mathcal{N}) \frac{|\mathcal{N}|}{A_n/4} \end{aligned} \quad (80)$$

since ν is uniformly distributed. Using that $|\mathcal{S}| < \eta n e^{m_2 \eta n}$,

$$|\mathcal{N}| \geq (A_n/4 - 2\eta n^2 e^{m_2 \eta n}),$$

hence from (80)

$$\begin{aligned} \mathbb{P}(\tau_n \geq A_n/2 | \nu \leq A_n/4) &\geq \mathbb{P}(\tau_n \geq A_n/2 | \nu \in \mathcal{N}) \left(1 - \frac{8\eta n^2 e^{m_2 \eta n}}{A_n} \right). \end{aligned} \quad (81)$$

Now, when $\nu \in \mathcal{N}$, all stopping times that could potentially depend on the transmitted codeword symbols are actually trivial, so the event $\{\tau_n \geq A_n/2\}$ is independent of the symbols sent at times $\nu, \nu+1, \dots, \nu+N-1$. Therefore,

$$\mathbb{P}(\tau_n \geq A_n/2 | \nu \in \mathcal{N}) = \mathbb{P}_*(\tau_n \geq A_n/2). \quad (82)$$

Combining (82) with (81) gives the desired claim (78).

Finally, to see that randomized stopping rules also cannot achieve asynchronism exponents larger than ηm , note that a randomized stopping rule can be viewed as simply a probability distribution over deterministic stopping rules. The previous analysis shows that for any deterministic stopping rule, and any asynchronism exponent larger than ηm , either the probability of error is large (e.g., at least $1/8$), or the expected delay is exponential in n . Therefore, the same holds for randomized stopping rules. ■

F. Comments on Error Criteria

We end this section by commenting on maximum versus average rate/error probability criteria. The results in this paper consider the rate defined with respect to maximum (over messages) reaction delay and consider maximum (over messages) error probability. Hence all the achievability results also hold when delay and error probability are averaged over messages.

To see that the converse results in this paper also hold for the average case, we use the following standard expurgation argument. Assume $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}$ is an (R, α) coding scheme

³⁰Any coding scheme that achieves vanishing error probability cannot have $\ln M$ grow faster than linearly with n , because of the limitation imposed by the capacity of the synchronous channel. Hence, if $\mathbb{E}(\tau_n - \nu)^+$ grows exponentially with n , the rate goes to zero exponentially with n .

where the error probability and the delay of $(\mathcal{C}_n, (\tau_n, \phi_n))$ are defined as

$$\epsilon_n \triangleq \frac{1}{M} \sum_{m=1}^M \mathbb{P}_m(\mathcal{E}),$$

and

$$\bar{\Delta}_n \triangleq \frac{1}{M} \sum_{m=1}^M \mathbb{E}_m(\tau_n - \nu)^+,$$

respectively. By definition of an (R, α) coding scheme, this means that given some arbitrarily small $\epsilon > 0$, and for all n large enough,

$$\epsilon_n \leq \epsilon$$

and

$$\frac{\ln M}{\bar{\Delta}_n} \geq R - \epsilon.$$

Hence, for n large enough and any $\delta > 1$, one can find a (nonzero) constant fraction of codewords $\mathcal{C}_n' \subset \mathcal{C}_n$ (\mathcal{C}_n' is the “expurgated” ensemble) that satisfies the following property: the rate defined with respect to maximum (over \mathcal{C}_n') delay is at least $(R - \epsilon)/\delta$ and the maximum error probability is less than $\eta\epsilon$, where $\eta = \eta(\delta) > 0$. One then applies the converse results to the expurgated ensemble to derive bounds on $(R/\delta, \alpha)$, and thus on (R, α) , since $\delta > 1$ can be chosen arbitrarily.

VI. CONCLUDING REMARKS

We analyzed a model for asynchronous communication which captures the situation when information is emitted infrequently. General upper and lower bounds on capacity were derived, which coincide in certain cases. The forms of these bounds are similar and have two parts: a mutual information part and a divergence part. The mutual information part is reminiscent of synchronous communication: to achieve a certain rate, there must be, on average, enough mutual information between the time information is sent and the time it is decoded. The divergence part is novel, and comes from asynchronism. Asynchronism introduces two additional error events that must be overcome by the decoder. The first event happens when the noise produces a channel output that looks as if it was generated by a codeword. The larger the level of asynchronism, the more likely this event becomes. The second event happens when the channel behaves atypically, which results in the decoder missing the codeword. When this event happens, the rate penalty is huge, on the order of the asynchronism level. As such, the second event contributes to increased average reaction delay, or equivalently, lowers the rate. The divergence part in our upper and lower bounds on capacity strikes a balance between these two events.

An important conclusion of our analysis is that, in general, training-based schemes are not optimal in the high rate, high asynchronism regime. In this regime, training-based architectures are unreliable, whereas it is still possible to achieve an arbitrarily low probability of error using strategies that combine synchronization with information transmission.

Finally, we note that further analysis is possible when we restrict attention to a simpler slotted communication model in which the possible transmission slots are nonoverlapping

and contiguous. In particular, for this more constrained model [?] develops a variety of results, among which is that except in somewhat pathological cases, training-based schemes are strictly suboptimal at all rates below the synchronous capacity. Additionally, the performance gap is quantified for the special cases of the binary symmetric and additive white Gaussian noise channels, where it is seen to be significant in the high rate regime but vanish in the limit of low rates. Whether the characteristics observed for the slotted model are also shared by unslotted models remains to be determined, and is a natural direction for future research.

ACKNOWLEDGMENTS

The authors are grateful to the reviewers for their insightful and detailed comments which very much contributed to improve the paper. The authors would also like to thank the associate editors Suhas Diggavi and Tsachy Weissman and the editor-in-chief Helmut Bölcskei for their care in handling this paper. This paper also benefited from useful discussions with Sae-Young Chung and Da Wang.

APPENDIX A PROOF OF REMARK 2 (P. 14)

To show that the random coding scheme proposed in the proof of Theorem 2 achieves (6) with equality, we show that

$$\alpha \leq \max_{P: I(PQ) \geq R} \min_{V \in \mathcal{P}^{\mathcal{Y}}} \max\{D(V\|(PQ)_{\mathcal{Y}}), D(V\|Q_{\star})\}. \quad (83)$$

Recall that, by symmetry of the encoding and decoding procedures, the average reaction delay is the same for any message. Hence

$$\Delta_n = \mathbb{E}_1(\tau_n - \nu)^+,$$

where \mathbb{E}_1 denotes expectation under the probability measure \mathbb{P}_1 , the channel output distribution when message 1 is sent, averaged over time and codebooks.

Suppose for the moment that

$$\mathbb{E}_1(\tau_n - \nu)^+ \geq n(1 - o(1)) \quad n \rightarrow \infty. \quad (84)$$

It then follows from Fano’s inequality that the input distribution P must satisfy $I(PQ) \geq R$. Hence, to establish (83) we will show that at least one of the following inequalities

$$\begin{aligned} D(V\|(PQ)_{\mathcal{Y}}) &\geq \alpha \\ D(V\|Q_{\star}) &\geq \alpha \end{aligned} \quad (85)$$

holds for any $V \in \mathcal{P}^{\mathcal{Y}}$. The arguments are similar to those used to establish Claim 3 of Theorem 3. Below we provide the key steps.

We proceed by contradiction and show that if both the inequalities in (85) are reversed, then the asymptotic rate is zero. To that aim we provide a lower bound on $\mathbb{E}_1(\tau_n - \nu)^+$.

Let τ_n' denote the time of the beginning of the decoding window, i.e., the first time when the previous n output symbols

have empirical distribution \hat{P} such that $D(\hat{P}||Q_*) \geq \alpha$. By definition, $\tau_n \geq \tau'_n$, so

$$\begin{aligned} \mathbb{E}_1(\tau_n - \nu)^+ &\geq \mathbb{E}_1(\tau'_n - \nu)^+ \\ &\geq \frac{1}{3} \sum_{t=1}^{A/3} \mathbb{P}_{1,t}(\tau'_n \geq 2A_n/3), \end{aligned} \quad (86)$$

where the second inequality follows from Markov's inequality, and where $\mathbb{P}_{1,t}$ denotes the probability measure at the output of the channel conditioned on the event that message 1 starts being sent at time t , and averaged over codebooks. Note that, because τ'_n is not a function of the codebook, there is no averaging on the stopping times.³¹

Fix $V \in \mathcal{P}_Y$. We lower bound each term $\mathbb{P}_{1,t}(\tau'_n \geq 2A_n/3)$ in the above sum as

$$\begin{aligned} \mathbb{P}_{1,t}(\tau'_n \geq 2A_n/3) &\geq \mathbb{P}_{1,t}(\tau'_n \geq 2A_n/3 | Y_t^{t+n-1} \in \mathcal{T}_V) \mathbb{P}_{1,t}(Y_t^{t+n-1} \in \mathcal{T}_V) \\ &\geq \mathbb{P}_{1,t}(\tau_n \geq 2A_n/3 | Y_t^{t+n-1} \in \mathcal{T}_V) e^{-nD_1} \text{poly}(n), \end{aligned} \quad (87)$$

where $D_1 \triangleq D(V||PQ)_Y$, and where the second inequality follows from Fact 2.

The key change of measure step (37) results now in the equality

$$\begin{aligned} \mathbb{P}_{1,t}(\tau'_n \geq 2A_n/3 | Y_t^{t+n-1} \in \mathcal{T}_V) &= \mathbb{P}_*(\tau'_n \geq 2A_n/3 | Y_t^{t+n-1} \in \mathcal{T}_V), \end{aligned} \quad (88)$$

which can easily be checked by noticing that the probability of any sequence y_t^{t+n-1} in \mathcal{T}_V is the same under $\mathbb{P}_{1,t}$. Substituting (88) into the right-hand side of (87), and using (86) and Fact 2, we get

$$\begin{aligned} \mathbb{E}_1(\tau_n - \nu)^+ &\geq e^{-n(D_1-D_2)} \text{poly}(n) \\ &\times \sum_{t=1}^{A/3} \mathbb{P}_*(\tau_n \geq 2A_n/3, Y_t^{t+n-1} \in \mathcal{T}_V), \end{aligned} \quad (89)$$

where $D_2 \triangleq D(V||Q_*)$. The rest of the proof consists in showing that if the two inequalities in (85) are reversed, then the right-hand side of the above inequality grows exponentially with n , which results in an asymptotic rate equal to zero. The arguments closely parallel the ones that prove Claim 3 of Theorem 3 (see from (38) onwards), and hence are omitted.

To conclude the proof we show (84). Using the alternate form of expectation for non-negative random variables $\mathbb{E}X = \sum_{k \geq 0} \mathbb{P}(X \geq k)$, we have

$$\begin{aligned} \mathbb{E}_1(\tau_n - \nu)^+ &\geq \sum_{i=1}^{g(n)} \mathbb{P}_1(\tau_n \geq \nu + i) \\ &\geq \sum_{i=1}^{g(n)} (1 - \mathbb{P}_1(\tau_n < \nu + i)) \\ &\geq g(n)(1 - \mathbb{P}_1(\tau_n \leq \nu + g(n))), \end{aligned}$$

³¹For different codebook realizations, stopping rule τ'_n is the same, by contrast with τ_n which depends on the codebook via the joint typicality criterion of the second phase.

where we defined

$$g(n) \triangleq n - \lceil n^{3/4} \rceil,$$

and where the last inequality follows from the fact that $\mathbb{P}_1(\tau_n < \nu + i)$ is a non-decreasing function of i . Since $g(n) = n(1 - o(1))$, to establish (84) it suffices to show that

$$\mathbb{P}_1(\tau_n \leq \nu + g(n)) = o(1) \quad (n \rightarrow \infty). \quad (90)$$

Since

$$\mathbb{P}_1(\tau_n < \nu) = o(1) \quad (n \rightarrow \infty),$$

as follows from computation steps in (22) and (23), to establish (90) it suffices to show that

$$\mathbb{P}_1(\nu \leq \tau_n \leq \nu + g(n)) = o(1) \quad (n \rightarrow \infty). \quad (91)$$

For $i \in \{0, 1, \dots, g(n)\}$ we have

$$\begin{aligned} \mathbb{P}_1(\tau_n = \nu + i) &\leq \mathbb{P}_1\left(\|\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} PQ\| \leq \mu \cdot |\mathcal{X}| \cdot |\mathcal{Y}|\right) \\ &= \sum_J \mathbb{P}_1\left(\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} = J\right) \end{aligned} \quad (92)$$

where the above summation is over all typical joint types, i.e., all $J \in \mathcal{P}_{\mathcal{X}, \mathcal{Y}}^{\mathcal{X}, \mathcal{Y}}$ such that

$$|\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}}(a, b) - J(a, b)| \leq \mu \quad (93)$$

for all $(a, b) \in \mathcal{X} \times \mathcal{Y}$.

We upper bound each term in this summation. First observe that event

$$\{\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} = J\},$$

for $i \in \{0, 1, \dots, g(n)\}$, involves random vector $Y_{\nu+i-n+1}^{\nu+i}$ which is partly generated by noise and partly generated by the transmitted codeword corresponding to message 1. In the following computation k refers to first symbols of $Y_{\nu+i-n+1}^{\nu+i}$ which are generated by noise, i.e., by definition $k = n - (i+1)$. Note that since $0 \leq i \leq g(n)$, we have

$$\lceil n^{3/4} \rceil - 1 \leq k \leq n - 1.$$

We have

$$\begin{aligned} \mathbb{P}_1(\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} = J) &= \sum_{\substack{J_1 \in \mathcal{P}_k \\ J_2 \in \mathcal{P}_{n-k} \\ kJ_1 + (n-k)J_2 = J}} \left(\sum_{(x^k, y^k): \hat{P}_{x^k, y^k} = J_1} P(x^k) Q_*(y^k) \right) \\ &\times \left(\sum_{(x^{n-k}, y^{n-k}): \hat{P}_{x^{n-k}, y^{n-k}} = J_2} \mathbb{P}(x^{n-k}, y^{n-k}) \right), \end{aligned} \quad (94)$$

where we used the following shorthand notations for probabilities

$$\begin{aligned} P(x^k) &\triangleq \prod_{j=1}^k P(x_j) \\ Q_*(y^k) &\triangleq \prod_{j=1}^k Q_*(y_j) \\ \mathbb{P}(x^{n-k}, y^{n-k}) &\triangleq \prod_{j=1}^k P(x_j) Q(y_j | x_j). \end{aligned}$$

Further, using Fact 2

$$\begin{aligned} \sum_{(x^k, y^k): \hat{P}_{x^k, y^k} = J_1} P(x^k) P_*(y^k) &= \sum_{x^k: \hat{P}_{x^k} = J_{1,x}} P(x^k) \sum_{y^k: \hat{P}_{y^k} = J_{1,y}} Q_*(y^k) \\ &\leq e^{-k(D(J_{1,x}||P) + D(J_{1,y}||Q_*))} \\ &\leq e^{-kD(J_{1,y}||Q_*)} \end{aligned} \quad (95)$$

where $J_{1,x}$ and $J_{1,y}$ denote the left and right marginals of J , respectively, and where the second inequality follows by non-negativity of divergence.

A similar calculation yields

$$\begin{aligned} \sum_{(x^{n-k}, y^{n-k}): \hat{P}_{x^{n-k}, y^{n-k}} = J_2} \mathbb{P}(x^{n-k}, y^{n-k}) &\leq e^{-(n-k)D(J_2||PQ)} \end{aligned} \quad (96)$$

From (94), (95), (96) and Fact 1 we get

$$\begin{aligned} \mathbb{P}_1(\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} = J) &\leq \text{poly}(n) \\ &\times \max_{\substack{J_1 \in \mathcal{P}_k^{\mathcal{X}, \mathcal{Y}} \\ J_2 \in \mathcal{P}_{n-k}^{\mathcal{X}, \mathcal{Y}} \\ kJ_1 + (n-k)J_2 = nJ \\ k: \lceil n^{3/4} \rceil - 1 \leq k \leq n-1}} \exp \left[-k(D(J_{1,y}||Q_*) \right. \\ &\quad \left. - (n-k)D(J_2||PQ)) \right]. \end{aligned} \quad (97)$$

The maximum on the right-hand side of (97) is equal to

$$\begin{aligned} \max_{\substack{J_1 \in \mathcal{P}_k^{\mathcal{X}, \mathcal{Y}} \\ J_2 \in \mathcal{P}_{n-k}^{\mathcal{X}, \mathcal{Y}} \\ kJ_1 + (n-k)J_2 = nJ \\ k: \lceil n^{3/4} \rceil - 1 \leq k \leq n-1}} \exp \left[-kD(J_1||Q_*) \right. \\ \left. - (n-k)D(J_2||PQ) \right]. \end{aligned} \quad (98)$$

We upper bound the argument of the above exponential via the log-sum inequality to get

$$\begin{aligned} &-kD(J_1||Q_*) - (n-k)D(J_2||PQ) \\ &\leq -nD(J_y||\delta Q_* + (1-\delta)(PQ)_y), \end{aligned} \quad (99)$$

where $\delta \triangleq k/n$. Using (99), we upper-bound expression (98) by

$$\begin{aligned} &\max_{\delta: n^{-1/4} - n^{-1} \leq \delta \leq 1} \exp \left[-nD(J_y||\delta Q_* + (1-\delta)(PQ)_y) \right] \\ &\leq \max_{\delta: n^{-1/4} - n^{-1} \leq \delta \leq 1} \exp \left[-n\Omega(\delta^2) \right] \\ &\leq \exp \left[-\Omega(n^{1/2}) \right], \end{aligned} \quad (100)$$

where for the first inequality we used Pinsker's inequality [7, Problem 17 p. 58]

$$D(P_1||P_2) \geq \frac{1}{2 \ln 2} \|P_1 - P_2\|^2,$$

and assume that μ is small enough and n is large enough for this inequality to be valid. Such μ and n exist whenever the distributions Q_* and $(PQ)_y$ are different.

It then follows from (97) that

$$\mathbb{P}_1(\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} = J) \leq \exp \left[-\Omega(n^{1/2}) \right],$$

hence, from (92) and Fact 1 we get

$$\mathbb{P}_1(\tau_n = \nu + i) \leq \exp \left[-\Omega(n^{1/2}) \right]$$

for $i \in \{0, 1, \dots, g(n)\}$. Finally a union bound over times yields the desired result (90) since $g(n) = O(n)$.

APPENDIX B PROOF OF THEOREM 5

The desired Theorem is a stronger version of [7, Corollary 1.9, p. 107], and its proof closely follows the proof of the latter.

Before proceeding, we recall the definitions of η -image and l -neighborhood of a set of sequences.

Definition 4 (η -image, [7] Definition 2.1.2 p. 101): A set $\mathcal{B} \subset \mathcal{Y}^n$ is an η -image of a set $\mathcal{A} \subset \mathcal{X}^n$ if $Q(\mathcal{B}|x) \geq \eta$ for all $x \in \mathcal{A}$. The minimum cardinality of η -images of \mathcal{A} is denoted $g_Q(\mathcal{A}, \eta)$.

Definition 5 (l -neighborhood, [7] p. 86): The l -neighborhood of a set $\mathcal{B} \subset \mathcal{Y}^n$ is the set

$$\Gamma^l \mathcal{B} \triangleq \{y^n \in \mathcal{Y}^n : d_H(\{y^n\}, \mathcal{B}) \leq l\}$$

where $d_H(\{y^n\}, \mathcal{B})$ denotes the Hamming distance between y^n and \mathcal{B} , i.e.,

$$d_H(\{y^n\}, \mathcal{B}) = \min_{\tilde{y}^n \in \mathcal{B}} d_H(y^n, \tilde{y}^n).$$

As other notation, for a given conditional probability $Q(y|x)$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and $x^n \in \mathcal{X}^n$, we define the set

$$\begin{aligned} \mathcal{T}_{[Q]}^n(x^n) &= \{y^n \in \mathcal{Y}^n : \\ &|\hat{P}_{x^n, y^n}(a, b) - \hat{P}_{x^n}(a)Q(b|a)| \leq \frac{1}{n^q}, \forall (a, b) \in \mathcal{X} \times \mathcal{Y}\} \end{aligned}$$

where $q \in (0, 1/2)$. To establish Theorem 5, we make use of the following three lemmas. Since we restrict attention to block coding schemes, i.e., coding scheme whose decoding happens at the fixed time n , we denote them simply by (\mathcal{C}_n, ϕ_n) instead of $(\mathcal{C}_n, (\gamma_n, \phi_n))$.

In the following, ϵ_n is always given by

$$\epsilon_n = (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n^{2q}/(2 \ln 2)).$$

Lemma 2: Given $\gamma \in (0, 1)$, $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, $P \in \mathcal{P}_n^{\mathcal{X}}$, and $\mathcal{A} \subset \mathcal{T}_P^n$, there exist (\mathcal{C}_n, ϕ_n) for each $n \geq n_o(\gamma, q, |\mathcal{X}|, |\mathcal{Y}|)$ such that

- 1) $c^n(m) \in \mathcal{A}$, for all $c^n(m) \in \mathcal{C}_n$
- 2) $\phi_n^{-1}(m) \subset \mathcal{T}_{[Q]}^n(c^n(m))$, $m \in \{1, 2, \dots, M\}$
- 3) the maximum error probability is upper bounded by $2\epsilon_n$
- 4) the rate satisfies

$$\frac{1}{n} \ln |\mathcal{C}_n| \geq \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon_n) - H(Q|P) - \gamma.$$

Proof of Lemma 2: The proof closely follows the proof of [7, Lemma 1.3, p. 101] since it essentially suffices to replace ϵ and γ in the proof of [7, Lemma 1.3, p. 101] with $2\epsilon_n$ and ϵ_n , respectively. We therefore omit the details here.

One of the steps of the proof consists in showing that

$$Q(\mathcal{T}_{[Q]}^n(x^n)|x^n) \geq 1 - \epsilon_n \quad (101)$$

for all $x^n \in \mathcal{X}^n$. To establish this, one proceeds as follow. Given $P \in \mathcal{P}_n^{\mathcal{X}}$ let \mathcal{D} denote the set of empirical conditional distributions $W(y|x) \in \mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}$ such that

$$|\hat{P}_{x^n}(a)W(b|a) - \hat{P}_{x^n}(a)Q(b|a)| \leq \frac{1}{n^q}$$

for all $(a, b) \in \mathcal{X} \times \mathcal{Y}$. We have

$$1 - Q(\mathcal{T}_{[Q]}^n(x^n)|x^n) = \sum_{W \in \mathcal{D}^c \cap \mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}} Q(\mathcal{T}_W^n(x^n)|x^n) \quad (102)$$

$$\leq \sum_{W \in \mathcal{D}^c \cap \mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}} e^{-nD(W\|Q|P)} \quad (103)$$

$$\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n \min_{W \in \mathcal{D}^c} D(W\|Q|P)) \quad (104)$$

$$\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n \min_{W \in \mathcal{D}^c} \|PW - PQ\|^2 / 2 \ln 2) \quad (105)$$

$$\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n^{2q}/2 \ln 2) \quad (106)$$

$$= \epsilon_n,$$

which shows (101). Inequality (103) follows from Fact 3, (104) follows from Fact 1, (105) follows from Pinsker's inequality (see, e.g., [7, Problem 17, p. 58]), and (106) follows from the definition of \mathcal{D} . ■

Lemma 3 ([7, Lemma 1.4, p. 104]): For every $\epsilon, \gamma \in (0, 1)$, if (\mathcal{C}_n, ϕ_n) achieves an error probability ϵ and $\mathcal{C}_n \subset \mathcal{T}_P^n$, then

$$\frac{1}{n} \ln |\mathcal{C}_n| < \frac{1}{n} \ln g_Q(\mathcal{C}_n, \epsilon + \gamma) - H(Q|P) + \gamma$$

whenever $n \geq n_o(|\mathcal{X}|, |\mathcal{Y}|, \gamma)$.

Since this lemma is established in [7, Lemma 1.4, p. 104], we omit its proof.

Lemma 4: For every $\gamma > 0$, $\epsilon \in (0, 1)$, $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, and $\mathcal{A} \subset \mathcal{X}^n$

$$\left| \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon) - \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon_n) \right| < \gamma$$

whenever $n \geq n_o(\gamma, q, |\mathcal{X}|, |\mathcal{Y}|)$.

Proof of Lemma 4: By the Blowing Up Lemma [7, Lemma 1.5.4, p. 92] and [7, Lemma 1.5.1, p. 86], given the sequence $\{\epsilon_n\}_{n \geq 1}$, there exist $\{l_n\}$ and $\{\eta_n\}$ such that $l_n/n \xrightarrow{n \rightarrow \infty} 0$ and $\eta_n \xrightarrow{n \rightarrow \infty} 1$, and such that the following two properties hold.

For any $\gamma > 0$ and $n \geq n_o(\gamma, q, |\mathcal{X}|, |\mathcal{Y}|)$

$$\frac{1}{n} \ln |\Gamma^{l_n} \mathcal{B}| - \frac{1}{n} \ln |\mathcal{B}| < \gamma \quad \text{for every } \mathcal{B} \subset \mathcal{Y}^n, \quad (107)$$

and for all $x^n \in \mathcal{X}^n$,

$$Q(\Gamma^{l_n} \mathcal{B}|x^n) \geq \eta_n \quad \text{whenever } Q(\mathcal{B}|x^n) \geq \epsilon_n. \quad (108)$$

Now, assuming that \mathcal{B} is an ϵ_n -image of \mathcal{A} with $|\mathcal{B}| = g_Q(\mathcal{A}, \epsilon_n)$, the relation (108) means that $\Gamma^{l_n} \mathcal{B}$ is an η_n -image of \mathcal{A} . Therefore we get

$$\begin{aligned} \frac{1}{n} \ln g_Q(\mathcal{A}, \eta_n) &\leq \frac{1}{n} \ln |\Gamma^{l_n} \mathcal{B}| \\ &\leq \gamma + \frac{1}{n} \ln |\mathcal{B}| \\ &= \gamma + \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon_n) \end{aligned} \quad (109)$$

where the second inequality follows from (107). Finally, since $\eta_n \rightarrow 1$ and $\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$, for n large enough we have

$$g_Q(\mathcal{A}, \epsilon) \leq g_Q(\mathcal{A}, \eta_n) \quad \text{and} \quad \epsilon_n \leq \epsilon,$$

and therefore from (109) we get

$$\frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon) \leq \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon_n) \leq \gamma + \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon)$$

yielding the desired result. ■

We now use these lemmas to establish Theorem 5. Choose $\epsilon, \gamma > 0$ such that $\epsilon + \gamma < l$. Let (\mathcal{C}_n, ϕ_n) be a coding scheme that achieves maximum error probability ϵ . Without loss of generality, we assume that $\mathcal{C}_n \subset \mathcal{T}_P^n$ (If not, group codewords into families of common type. The largest family of codewords has error probability no larger than ϵ , and its rate is essentially the same as the rate of the original code \mathcal{C}_n .) Therefore

$$\begin{aligned} \frac{1}{n} \ln |\mathcal{C}_n| &\leq \frac{1}{n} \ln g_Q(\mathcal{C}_n, \epsilon + \gamma) - H(Q|P) + \gamma \\ &\leq \frac{1}{n} \ln g_Q(\mathcal{C}_n, l) - H(Q|P) + \gamma \\ &\leq \frac{1}{n} \ln g_Q(\mathcal{C}_n, \epsilon_n) - H(Q|P) + 2\gamma \end{aligned} \quad (110)$$

for $n \geq n_o(\gamma, l, |\mathcal{X}|, |\mathcal{Y}|)$, where the first and third inequalities follow from Lemmas 3 and 4, respectively, and where the second inequality follows since $g_Q(\mathcal{C}_n, \epsilon)$ is nondecreasing in ϵ . On the other hand, by Lemma 2, there exists a coding scheme $(\mathcal{C}'_n, \phi'_n)$, with $\mathcal{C}'_n \subset \mathcal{C}_n$ that achieves a probability of error upper bounded by $2\epsilon_n$ and such that its rate satisfies

$$\frac{1}{n} \ln |\mathcal{C}'_n| \geq \frac{1}{n} \ln g_Q(\mathcal{C}_n, \epsilon_n) - H(Q|P) - \gamma \quad (111)$$

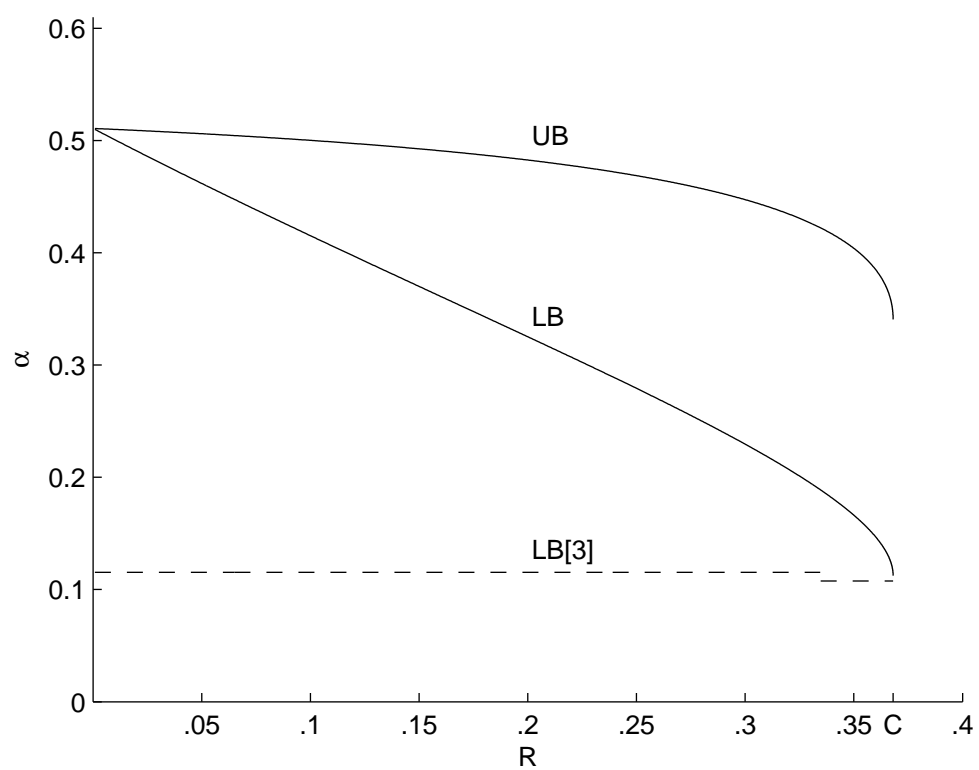
for $n \geq n_o(\gamma, q, |\mathcal{X}|, |\mathcal{Y}|)$. From (110) and (111) we deduce the rate of \mathcal{C}'_n is lower bounded as

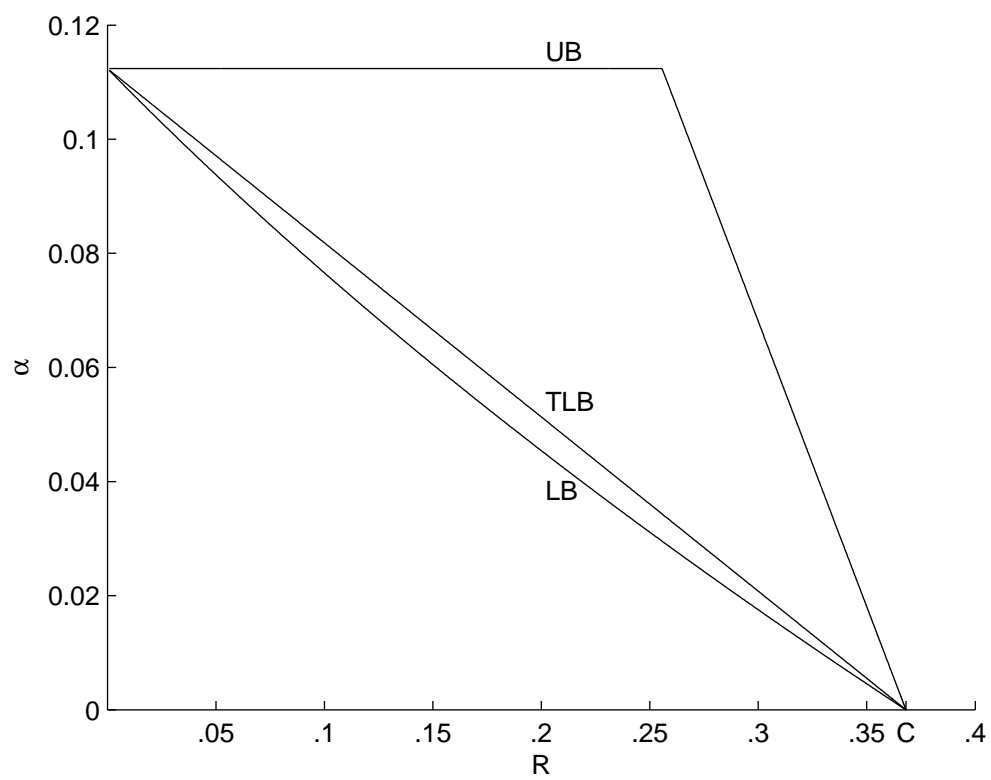
$$\frac{1}{n} \ln |\mathcal{C}'_n| \geq \frac{1}{n} \ln |\mathcal{C}_n| - 3\gamma$$

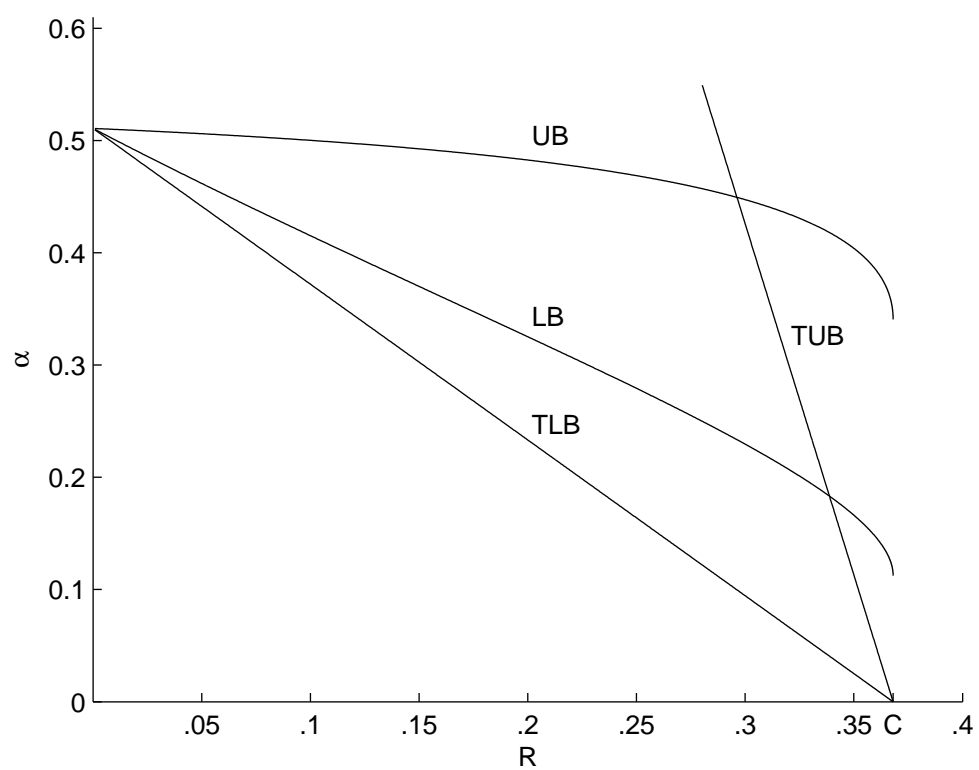
whenever $n \geq n_o(\gamma, l, q, |\mathcal{X}|, |\mathcal{Y}|)$. This yields the desired result. ■

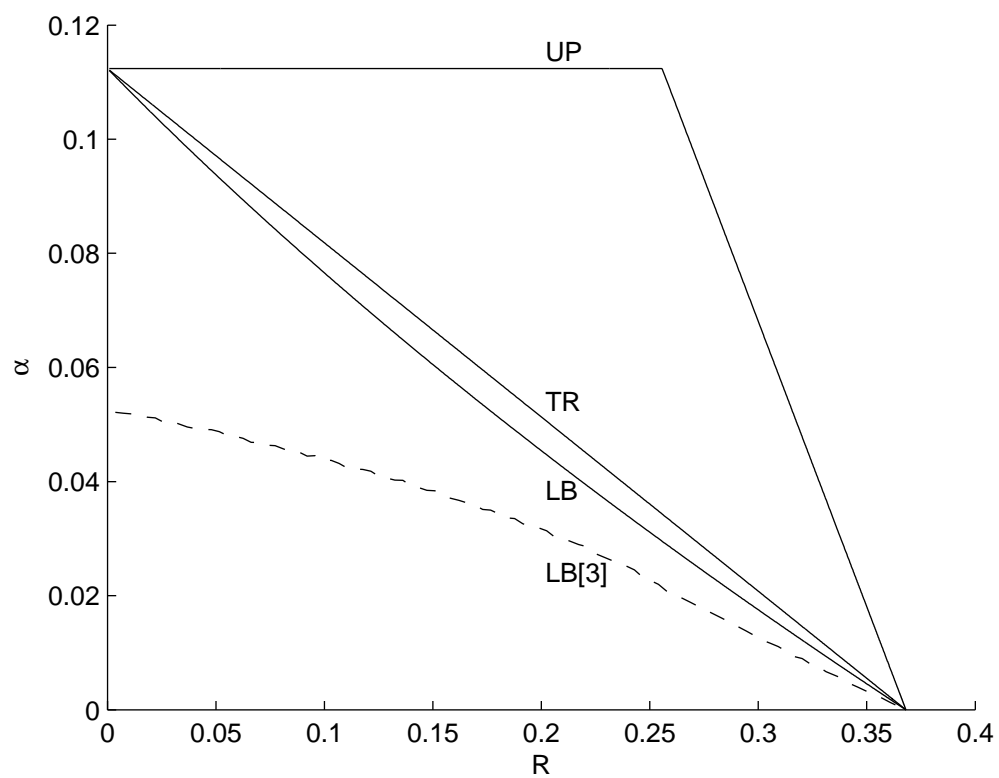
REFERENCES

- [1] A. Tchamkerten, V. Chandar, and G. Wornell, "On the capacity region of asynchronous channels," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2008.
- [2] V. Chandar, A. Tchamkerten, and G. Wornell, "Training-based schemes are suboptimal for high rate asynchronous communication," in *Proc. IEEE Information Theory Work. (ITW)*, Taormina, October 2009.
- [3] A. Tchamkerten, V. Chandar, and G. Wornell, "Communication under strong asynchronism," *IEEE Trans. Inform. Th.*, vol. 55, no. 10, pp. 4508–4528, October 2009.
- [4] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423, October 1948.
- [5] V. Chandar, A. Tchamkerten, and D. Tse, "Asynchronous capacity per unit cost," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, june 2010, pp. 280 –284.
- [6] —, "Asynchronous capacity per unit cost," *CoRR*, vol. abs/1007.4872, 2010.
- [7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels*. New York: Academic Press, 1981.
- [8] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, 2nd edition*. MIT Press, McGraw-Hill Book Company, 2000.
- [9] V. Chandar, A. Tchamkerten, and G. Wornell, "Optimal sequential frame synchronization," *IEEE Trans. Inform. Th.*, vol. 54, no. 8, pp. 3725–3728, 2008.
- [10] I. Csiszár and P. Narayan, "Arbitrarily varying channels with constrained inputs and states," *IEEE Transactions on Information Theory*, vol. 34, no. 1, pp. 27–34, 1988.
- [11] T. Cover and J. Thomas, *Elements of information theory*. New York: Wiley, 2006.
- [12] R. G. Gallager, *Information Theory and Reliable Communication*. Budapest: Wiley, 1968.









Asynchronous Communication: Capacity Bounds and Suboptimality of Training

Aslan Tchamkerten, Venkat Chandar, and Gregory W. Wornell *Fellow, IEEE*

Abstract—Several aspects of the problem of asynchronous point-to-point communication without feedback are developed when the source is highly intermittent. In the system model of interest, the codeword is transmitted at a random time within a prescribed window whose length corresponds to the level of asynchronism between the transmitter and the receiver. The decoder operates sequentially and communication rate is defined as the ratio between the message size and the elapsed time between when transmission commences and when the decoder makes a decision.

For such systems, general upper and lower bounds on capacity as a function of the level of asynchronism are established, and are shown to coincide in some nontrivial cases. From these bounds, several properties of this asynchronous capacity are derived. In addition, the performance of training-based schemes is investigated. It is shown that such schemes, which implement synchronization and information transmission on separate degrees of freedom in the encoding, cannot achieve the asynchronous capacity in general, and that the penalty is particularly significant in the high-rate regime.

Index Terms—asynchronous communication; bursty communication; error exponents; sequential decoding; sparse communication; synchronization

I. INTRODUCTION

INFORMATION-THEORETIC analysis of communication systems frequently ignores synchronization issues. In many applications where large amounts of data are to be transmitted, such simplifications may be justified. Simply prepending a suitable synchronization preamble to the initial data incurs negligible overhead yet ensures that the transmitter and the receiver are synchronized. In turn, various coding techniques (e.g., graph based codes, polar codes) may guarantee delay optimal communication for data transmission in the sense that they can achieve the capacity of the synchronous channel.

In quantifying the impact due to a lack of synchronization between a transmitter and a receiver, it is important to note that asynchronism is a relative notion that depends on the size of the data to be transmitted. For instance, in the above “low

asynchronism” setting it is implicitly assumed that the data is large with respect to the timing uncertainty.

In a growing number of applications, such as many involving sensor networks, data is transmitted in a bursty manner. An example would be a sensor in a monitoring system. By contrast with the previous setting, here timing uncertainty is large with respect to the data to be transmitted.

To communicate in such “high asynchronism” regimes, one can use the traditional preamble based communication scheme for each block. Alternatively, one can pursue a fundamentally different strategy in which synchronization is integrated into the encoding of the data, rather than separated from it.

To evaluate the relative merits of such diverse strategies, and more generally to explore fundamental performance limits, we recently introduced a general information-theoretic model for asynchronous communication in [3]. This model extends Shannon’s original communication model [4] to include asynchronism. In this model, the message is encoded into a codeword of fixed length, and this codeword starts being sent across a discrete memoryless channel at a time instant that is randomly and uniformly distributed over some predefined transmission window. The size of this window is known to transmitter and receiver, and the level of asynchronism in the system is governed by the size of the window with respect to the codeword length. Outside the information transmission period, whose duration equals the codeword length, the transmitter remains idle and the receiver observes noise, i.e., random output symbols. The receiver uses a sequential decoder whose scope is twofold: decide when to decode and what message to declare.

The performance measure is the communication rate which is defined as the ratio between the message size and the average delay between when transmission starts and when the message is decoded. Capacity is the supremum of achievable rates, i.e., rates for which vanishing error probability can be guaranteed in the limit of long codeword length.

The scaling between the transmission window and the codeword length that meaningfully quantifies the level of asynchronism in the system turns out to be exponential, i.e., $A = e^{\alpha n}$ where A denotes the size of the transmission window, where n denotes the codeword length, and where α denotes the asynchronism exponent. Indeed, as discussed in [3], if A scales subexponentially in n , then asynchronism doesn’t impact communication: the asynchronous capacity is equal to the capacity of the synchronous channel. By contrast, if the window size scales superexponentially, then the asynchrony is generally catastrophic. Hence, exponential asynchronism is the interesting regime and we aim to compute

This work was supported in part by an Excellence Chair Grant from the French National Research Agency (ACE project), and by the National Science Foundation under Grant No. CCF-1017772. This work was presented in part at the IEEE International Symposium on Information Theory, Toronto, Canada, July 2008 [1], and at the IEEE Information Theory Workshop, Taormina, Italy, October 2009 [2].

A. Tchamkerten is with the Department of Communications and Electronics, Telecom ParisTech, 75634 Paris Cedex 13, France. (Email: aslan.tchamkerten@telecom-paristech.fr).

V. Chandar is with MIT Lincoln Laboratory, Lexington, MA 02420 (Email: vchandar@mit.edu).

G. W. Wornell is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 (Email: gww@mit.edu).

capacity as a function of the asynchronism exponent.

For further motivation and background on the model, including a summary of related models (e.g., the insertion, deletion, and substitution channel model, and the detection and isolation model) we refer to [3, Section II]. Accordingly, we omit such material from the present paper.

The first main result in [3] is the characterization of the synchronization threshold, which is defined as the largest asynchronism exponent for which it is still possible to guarantee reliable communication—this result is recalled in Theorem 1 of Section IV.

The second main result in [3] (see [3, Theorem 1]) is a lower bound to capacity. A main consequence of this bound is that for any rate below the capacity of the synchronous channel it is possible to accommodate a non-trivial asynchronism level, i.e., a positive asynchronism exponent.

While this work focuses on rate, an alternative performance metric is the minimum energy (or, more generally, the minimum cost) needed to transmit one bit of information asynchronously. For this metric, [5], [6] establishes the capacity per unit cost for the above bursty communication setup.

We now provide a brief summary of the results contained in this paper:

- *General capacity lower bound, Theorems 2 and 1.* Theorem 2 provides a lower bound to capacity which is obtained by considering a coding scheme that performs synchronization and information transmission jointly. The derived bound results in a much simpler and often much better lower bound than the one obtained in [3, Theorem 1]. Theorem 2, which holds for arbitrary discrete memoryless channels, also holds for a natural Gaussian setting, which yields Theorem 1.
- *General capacity upper bound, Theorem 3.* This bound and the above lower bound, although not tight in general, provide interesting and surprising insights into the asynchronous capacity. For instance, Corollary 2 says that, in general, it is possible to reliably achieve a communication rate equal to the capacity of the synchronous channel while operating at a strictly positive asynchronism exponent. In other words, it is possible to accommodate both a high rate and an exponential asynchronism. Another insight is provided by Corollary 3, which relates to the very low rate communication regime. This result says that, in general, one needs to (sometimes significantly) back off from the synchronization threshold in order to be able to accommodate a positive rate. As a consequence, capacity as a function of the asynchronism exponent does not, in general, strictly increase as the latter decreases.
- *Capacity for channels with infinite synchronization threshold, Theorem 4.* For the class of channels for which there exists a particular channel input which can't be confused with noise, a closed-form expression for capacity is established.
- *Suboptimality of training based schemes, Theorem 6, Corollaries 4 and 5.* These results show that communication strategies that separate synchronization from information transmission do not achieve the asynchronous

capacity in general.

- *Good synchronous codes, Theorem 5.* This result may be of independent interest and relates to synchronous communication. It says that any codebook that achieves a nontrivial error probability contains a large subcodebook, whose rate is almost the same as the rate of the original codebook, and whose error probability decays exponentially with the blocklength with a suitable decoder. This result, which is a byproduct of our analysis, is a stronger version of [7, Corollary 1.9, p. 107] and its proof amounts to a tightening of some of the arguments in the proof of the latter.

It is worth noting that most of our proof techniques differ in some significant respects from more traditional capacity analysis for synchronous communication—for example, we make little use of Fano's inequality for converse arguments. The reason for this is that there are decoding error events specific to asynchronous communication. One such event is when the decoder, unaware of the information transmission time, declares a message before transmission even starts.

An outline of the paper is as follows. Section II summarizes some notational conventions and standard results we make use of throughout the paper. Section III describes the communication model of interest. Section IV contains our main results, and Section V is devoted to the proofs. Section VI contains some concluding remarks.

II. NOTATION AND PRELIMINARIES

In general, we reserve upper case letters for random variables (e.g., X) and lower case letters to denote their corresponding sample values (e.g., x), though as is customary, we make a variety of exceptions. Any potential confusion is generally avoided by context. In addition, we use x_i^j to denote the sequence x_i, x_{i+1}, \dots, x_j , for $i \leq j$. Moreover, when $i = 1$ we use the usual simpler notation x^n as an alternative to x_1^n . Additionally, \triangleq denotes “equality by definition.”

Events (e.g., \mathcal{E}) and sets (e.g., \mathcal{S}) are denoted using calligraphic fonts, and if \mathcal{E} represents an event, \mathcal{E}^c denotes its complement. As additional notation, $\mathbb{P}[\cdot]$ and $\mathbb{E}[\cdot]$ denote the probability and expectation of their arguments, respectively, $\|\cdot\|$ denotes the L_1 norm of its argument, $|\cdot|$ denotes absolute value if its argument is numeric, or cardinality if its argument is a set, $\lfloor \cdot \rfloor$ denotes the integer part of its argument, $a \wedge b \triangleq \min\{a, b\}$, and $x^+ \triangleq \max\{0, x\}$. Furthermore, we use \subset to denote nonstrict set inclusion, and use the Kronecker notation $\mathbb{1}(\mathcal{A})$ for the function that takes value one if the event \mathcal{A} is true and zero otherwise.

We also make use of some familiar order notation for asymptotics (see, e.g., [8, Chapter 3]). We use $o(\cdot)$ and $\omega(\cdot)$ to denote (positive or negative) quantities that grow strictly slower and strictly faster, respectively, than their arguments; e.g., $o(1)$ denotes a vanishing term and $n/\ln n = \omega(\sqrt{n})$. We also use $O(\cdot)$ and $\Omega(\cdot)$, defined analogously to $o(\cdot)$ and $\omega(\cdot)$, respectively, but without the strictness constraint. Finally, we use $\text{poly}(\cdot)$ to denote a function that does not grow or decay faster than polynomially in its argument.

We use $\mathbb{P}(\cdot)$ to denote the probability of its argument, and use $\mathcal{P}^{\mathcal{X}}$, $\mathcal{P}^{\mathcal{Y}}$, and $\mathcal{P}^{\mathcal{X}, \mathcal{Y}}$ to denote the set of distributions over

the finite alphabets \mathcal{X} , \mathcal{Y} , and $\mathcal{X} \times \mathcal{Y}$ respectively, and use $\mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ to denote the set of conditional distributions of the form $V(y|x)$ for $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

For a memoryless channel characterized by channel law $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, the probability of the output sequence $y^n \in \mathcal{Y}^n$ given an input sequence $x^n \in \mathcal{X}^n$ is

$$Q(y^n|x^n) \triangleq \prod_{i=1}^n Q(y_i|x_i).$$

Throughout the paper, Q always refers to the underlying channel and C denotes its synchronous capacity.

Additionally, we use J_X and J_Y to denote the left and right marginals, respectively, of the joint distribution $J \in \mathcal{P}^{\mathcal{X}, \mathcal{Y}}$, i.e.,

$$J_X(x) \triangleq \sum_{y \in \mathcal{Y}} J(x, y) \quad \text{and} \quad J_Y(y) \triangleq \sum_{x \in \mathcal{X}} J(x, y).$$

We define all information measures relative to the natural logarithm. Thus, the entropy associated with $P \in \mathcal{P}^{\mathcal{X}}$ is¹

$$H(P) \triangleq - \sum_{x \in \mathcal{X}} P(x) \ln P(x),$$

and the conditional entropy associated with $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ and $P \in \mathcal{P}^{\mathcal{X}}$ is

$$H(Q|P) \triangleq - \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} Q(y|x) \ln Q(y|x).$$

Similarly, the mutual information induced by $J(\cdot, \cdot) \in \mathcal{P}^{\mathcal{X}, \mathcal{Y}}$ is

$$I(J) \triangleq \sum_{(x, y) \in \mathcal{X} \times \mathcal{Y}} J(x, y) \ln \frac{J(x, y)}{J_X(x)J_Y(y)},$$

so

$$I(PQ) \triangleq \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} Q(y|x) \ln \frac{Q(y|x)}{(PQ)_Y(y)}$$

for $P \in \mathcal{P}^{\mathcal{X}}$ and $W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$. Furthermore, the information divergence (Kullback-Leibler distance) between $P_1 \in \mathcal{P}^{\mathcal{X}}$ and $P_2 \in \mathcal{P}^{\mathcal{X}}$ is

$$D(P_1 \| P_2) \triangleq \sum_{x \in \mathcal{X}} P_1(x) \ln \frac{P_1(x)}{P_2(x)},$$

and conditional information divergence is denoted using

$$\begin{aligned} D(W_1 \| W_2 | P) &\triangleq \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} W_1(y|x) \ln \frac{W_1(y|x)}{W_2(y|x)} \\ &\triangleq D(PW_1 \| PW_2), \end{aligned}$$

where $P \in \mathcal{P}^{\mathcal{X}}$ and $W_1, W_2 \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$. As a specialized notation, we use

$$D_B(\epsilon_1 \| \epsilon_2) \triangleq \epsilon_1 \ln \left(\frac{\epsilon_1}{\epsilon_2} \right) + (1 - \epsilon_1) \ln \left(\frac{1 - \epsilon_1}{1 - \epsilon_2} \right)$$

to denote the divergence between Bernoulli distributions with parameters $\epsilon_1, \epsilon_2 \in [0, 1]$.

¹In the definition of all such information measures, we use the usual convention $0 \ln(0/0) = 0$.

We make frequent use of the method of types [7, Chapter 1.2]. In particular, \hat{P}_{x^n} denotes the empirical distribution (or type) of a sequence $x^n \in \mathcal{X}^n$, i.e.,²

$$\hat{P}_{x^n}(x) \triangleq \frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i = x).$$

The joint empirical distribution $\hat{P}_{(x^n, y^n)}$ for a sequence pair (x^n, y^n) is defined analogously, i.e.,

$$\hat{P}_{x^n, y^n}(x, y) \triangleq \frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i = x, y_i = y),$$

and, in turn, a sequence y^n is said to have a conditional empirical distribution $\hat{P}_{y^n|x^n} \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ given x^n if for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\hat{P}_{x^n, y^n}(x, y) \triangleq \hat{P}_{x^n}(x) \hat{P}_{y^n|x^n}(y|x).$$

As additional notation, $P \in \mathcal{P}^{\mathcal{X}}$ is said to be an n -type if $nP(x)$ is an integer for all $x \in \mathcal{X}$. The set of all n -types over an alphabet \mathcal{X} is denoted using $\mathcal{P}_n^{\mathcal{X}}$. The n -type class of P , denoted using \mathcal{T}_P^n , is the set of all sequences x^n that have type P , i.e., such that $\hat{P}_{x^n} = P$. A set of sequences is said to have constant composition if they belong to the same type class. When clear from the context, we sometimes omit the superscript n and simply write \mathcal{T}_P . For distributions on the alphabet $\mathcal{X} \times \mathcal{Y}$ the set of joint n -types $\mathcal{P}_n^{\mathcal{X}, \mathcal{Y}}$ is defined analogously. The set of sequences y^n that have a conditional type W given x^n is denoted by $\mathcal{T}_W(x^n)$, and $\mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}$ denotes the set of empirical conditional distributions, i.e., the set of $W \in \mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}$ such that $W = \hat{P}_{y^n|x^n}(y|x)$ for some $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$.

Finally, the following three standard type results are often used in our analysis.

Fact 1 ([7, Lemma 1.2.2]):

$$\begin{aligned} |\mathcal{P}_n^{\mathcal{X}}| &\leq (n+1)^{|\mathcal{X}|} \\ |\mathcal{P}_n^{\mathcal{X}, \mathcal{Y}}| &\leq (n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|} \\ |\mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}| &\leq (n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|}. \end{aligned}$$

Fact 2 ([7, Lemma 1.2.6]): If X^n is independent and identically distributed (i.i.d.) according to $P_1 \in \mathcal{P}^{\mathcal{X}}$, then

$$\frac{1}{(n+1)^{|\mathcal{X}|}} e^{-nD(P_2 \| P_1)} \leq \mathbb{P}(X^n \in \mathcal{T}_{P_2}) \leq e^{-nD(P_2 \| P_1)},$$

for any $P_2 \in \mathcal{P}_n^{\mathcal{X}}$.

Fact 3 ([7, Lemma 1.2.6]): If the input $x^n \in \mathcal{X}^n$ to a memoryless channel $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ has type $P \in \mathcal{P}^{\mathcal{X}}$, then the probability of observing a channel output sequence Y^n which lies in $\mathcal{T}_W(x^n)$ satisfies

$$\begin{aligned} \frac{1}{(n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|}} e^{-nD(W \| Q|P)} &\leq \mathbb{P}(Y^n \in \mathcal{T}_W(x^n) | x^n) \\ &\leq e^{-nD(W \| Q|P)} \end{aligned}$$

for any $W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ such that $\mathcal{T}_W(x^n)$ is non-empty.

²When the sequence that induces the empirical type is clear from context, we omit the subscript and write simply \hat{P} .

III. MODEL AND PERFORMANCE CRITERION

The asynchronous communication model of interest captures the setting where infrequent delay-sensitive data must be reliably communicated. For a discussion of this model and its connections with related communication and statistical models we refer to [3, Section II].

We consider discrete-time communication without feedback over a discrete memoryless channel characterized by its finite input and output alphabets \mathcal{X} and \mathcal{Y} , respectively, and transition probability matrix $Q(y|x)$, for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. Without loss of generality, we assume that for all $y \in \mathcal{Y}$ there is some $x \in \mathcal{X}$ for which $Q(y|x) > 0$.

There are $M \geq 2$ messages $m \in \{1, 2, \dots, M\}$. For each message m , there is an associated codeword

$$c^n(m) \triangleq c_1(m) c_2(m) \cdots c_n(m),$$

which is a string of n symbols drawn from \mathcal{X} . The M codewords form a codebook \mathcal{C}_n (whence $|\mathcal{C}_n| = M$). Communication takes place as follows. The transmitter selects a message m randomly and uniformly over the message set and starts sending the corresponding codeword $c^n(m)$ at a random time ν , unknown to the receiver, independent of $c^n(m)$, and uniformly distributed over $\{1, 2, \dots, A\}$, where $A \triangleq e^{n\alpha}$ is referred to as the *asynchronism level* of the channel, with α termed the associated *asynchronism exponent*. The transmitter and the receiver know the integer parameter $A \geq 1$. The special case $A = 1$ (i.e., $\alpha = 0$) corresponds to the classical synchronous communication scenario.

When a codeword is transmitted, a noise-corrupted version of the codeword is obtained at the receiver. When the transmitter is silent, the receiver observes only noise. To characterize the output distribution when no input is provided to the channel, we make use of a specially designated “no-input” symbol \star in the input alphabet \mathcal{X} , as depicted in Figs. 1 and 2. Specifically,

$$Q_\star \triangleq Q(\cdot|\star) \quad (1)$$

characterizes the noise distribution of the channel. Hence, conditioned on the value of ν and on the message m to be conveyed, the receiver observes independent symbols $Y_1, Y_2, \dots, Y_{A+n-1}$ distributed as follows. If

$$t \in \{1, 2, \dots, \nu - 1\}$$

or

$$t \in [\nu + n, \nu + n + 1, \dots, A + n - 1],$$

the distribution of Y_t is Q_\star . If

$$t \in \{\nu, \nu + 1, \dots, \nu + n - 1\},$$

the distribution of Y_t is $Q(\cdot|c_{t-\nu+1}(m))$. Note that since the transmitter can choose to be silent for arbitrary portions of its length- n transmission as part of its message-encoding strategy, the symbol \star is eligible for use in the codebook design.

The decoder takes the form of a sequential test (τ, ϕ) , where τ is a stopping time, bounded by $A + n - 1$, with respect to the output sequence Y_1, Y_2, \dots , indicating when decoding happens, and where ϕ denotes a decision rule that declares the decoded message; see Fig. 2. Recall that a stopping time

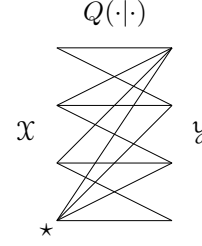


Fig. 1. Graphical depiction of the transmission matrix for an asynchronous discrete memoryless channel. The “no input” symbol \star is used to characterize the channel output when the transmitter is silent.

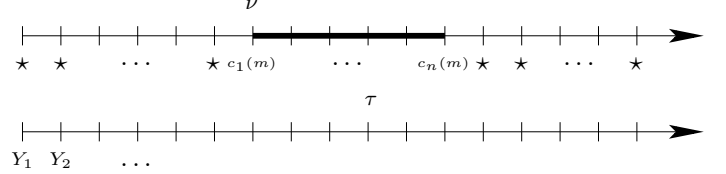


Fig. 2. Temporal representation of the channel input sequence (upper axis) and channel output sequence (lower axis). At time ν message m starts being sent and decoding occurs at time τ . Since ν is unknown at the receiver, the decoding time may be before the entire codeword has been received, potentially (but not necessarily) resulting in a decoding error.

τ (deterministic or randomized) is an integer-valued random variable with respect to a sequence of random variables $\{Y_i\}_{i=1}^\infty$ so that the event $\{\tau = t\}$, conditioned on $\{Y_i\}_{i=1}^t$, is independent of $\{Y_i\}_{i=t+1}^\infty$, for all $t \geq 1$. The function ϕ is then defined as any \mathcal{F}_τ -measurable map taking values in $\{1, 2, \dots, M\}$, where $\mathcal{F}_1, \mathcal{F}_2, \dots$ is the natural filtration induced by the process Y_1, Y_2, \dots .

A code is an encoder/decoder pair $(\mathcal{C}, (\tau, \phi))$.³

The performance of a code operating over an asynchronous channel is quantified as follows. First, we define the maximum (over messages), time-averaged decoding error probability⁴

$$\mathbb{P}(\mathcal{E}) = \max_m \frac{1}{A} \sum_{t=1}^A \mathbb{P}_{m,t}(\mathcal{E}), \quad (2)$$

where \mathcal{E} indicates the event that the decoded message does not correspond to the sent message, and where the subscripts m, t indicate the conditioning on the event that message m starts being sent at time $\nu = t$. Note that by definition we have

$$\mathbb{P}_{m,t}(\mathcal{E}) = \mathbb{P}_{m,t}(\phi(Y^\tau) \neq m).$$

Second, we define communication rate with respect to the average elapsed time between the time the codeword starts being sent and the time the decoder makes a decision, i.e.,

$$R = \frac{\ln M}{\Delta}, \quad (3)$$

where

$$\Delta = \max_m \frac{1}{A} \sum_{t=1}^A \mathbb{E}_{m,t}(\tau - t)^+, \quad (4)$$

³Note that the proposed asynchronous discrete-time communication model still assumes some degree of synchronization since transmitter and receiver are supposed to have access to clocks ticking at unison. This is sometimes referred to as frame asynchronous symbol synchronous communication.

⁴Note that there is a small abuse of notation as $\mathbb{P}(\mathcal{E})$ need not be a probability.

where x^+ denotes $\max\{0, x\}$, and where $\mathbb{E}_{m,t}$ denotes the expectation with respect to $\mathbb{P}_{m,t}$.⁵

With these definitions, the class of communication strategies of interest is as follows.

Definition 1 (*(R, α) Coding Scheme*): A pair (R, α) with $R \geq 0$ and $\alpha \geq 0$ is achievable if there exists a sequence $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ of codes, indexed by the codebook length n , that asymptotically achieves a rate R at an asynchronism exponent α . This means that for any $\epsilon > 0$ and every n large enough, the code $(\mathcal{C}_n, (\tau_n, \phi_n))$

- 1) operates under asynchronism level $A_n = e^{(\alpha - \epsilon)n}$;
- 2) yields a rate at least equal to $R - \epsilon$;
- 3) achieves a maximum error probability of at most ϵ .

An (R, α) coding scheme is a sequence $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ that achieves the rate-exponent pair (R, α) .

In turn, capacity for our model is defined as follows.

Definition 2 (*Asynchronous Capacity*): For given $\alpha \geq 0$, the asynchronous capacity $R(\alpha)$ is the supremum of the set of rates that are achievable at asynchronism exponent α . Equivalently, the asynchronous capacity is characterized by $\alpha(R)$, defined as the supremum of the set of asynchronism exponents that are achievable at rate $R \geq 0$.

Accordingly, we use the term ‘‘asynchronous capacity’’ to designate either $R(\alpha)$ or $\alpha(R)$. While $R(\alpha)$ may have the more natural immediate interpretation, most of our results are more conveniently expressed in terms of $\alpha(R)$.

In agreement with our notational convention, the capacity of the synchronous channel, which corresponds to the case where $\alpha = 0$, is simply denoted by C instead of $R(0)$. Throughout the paper we only consider channels with $C > 0$.

Remark 1: One could alternatively consider the rate with respect to the duration the transmitter occupies the channel and define it with respect to the block length n . In this case capacity is a special case of the general asynchronous capacity per unit cost result [5, Theorem 1].

In [3], [9] it is shown that reliable communication is possible if and only if the asynchronism exponent α does not exceed a limit referred to as the ‘‘synchronization threshold.’’

Theorem 1 ([3, Theorem 2], [9]): If the asynchronism exponent is strictly smaller than the *synchronization threshold*

$$\alpha_o \triangleq \max_x D(Q(\cdot|x) \| Q_\star) = \alpha(R = 0),$$

then there exists a coding scheme $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ that achieves a maximum error probability tending to zero as $n \rightarrow \infty$.

Conversely, any coding scheme $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ that operates at an asynchronism exponent strictly greater than the synchronization threshold, achieves (as $n \rightarrow \infty$) a maximum probability of error equal to one.

Moreover,⁶

$$\alpha_o > 0 \quad \text{if and only if} \quad C > 0.$$

A few comments are in order. The cause of unreliable communication above the synchronization threshold is the following. When asynchronism is so large, with probability

approaching one pure noise mimics a codeword for *any* codebook (regardless of the rate) before the actual codeword even starts being sent.⁷ This results in an error probability of at least $1/2$ since, by our model assumption, the message set contains at least two messages. On the other hand, below the synchronization threshold reliable communication is possible. If the codebook is properly chosen, the noise won’t mimic any codeword with probability tending to one, which allows the decoder to reliably detect the sent message.

Note that

$$\alpha_o = \infty$$

if and only if pure noise can’t generate all channel outputs, i.e., if and only if $Q_\star(y) = 0$ for some $y \in \mathcal{Y}$. Indeed, in this case it is possible to avoid the previously mentioned decoding confusion by designing codewords (partly) composed of symbols that generate channel outputs which are impossible to generate with pure noise.

The last claim in Theorem 1 says that reliable asynchronous communication is possible if and only if reliable synchronous communication is possible. That the former implies the latter is obvious since asynchronism can only hurt communication. That the latter implies the former is perhaps less obvious, and a high-level justification is as follows. When $C > 0$, at least two channel inputs yield different conditional output distributions, for otherwise the input-output mutual information is zero regardless of the input distribution. Hence, $Q(\cdot|\star) \neq Q(\cdot|x)$ for some $x \neq \star$. Now, by designing codewords mainly composed of x it is possible to reliably signal the codeword’s location to the decoder even under an exponential asynchronism, since the channel outputs look statistically different than noise during the message transmission. Moreover, if the message set is small enough, it is possible to guarantee reliable message location and successfully identify which message from the message set was sent. Therefore, exponential asynchronism can be accommodated, hence $\alpha_o > 0$.

Finally, it should be pointed out that in [3] all the results are stated with respect to average (over messages) delay and error probability in place of maximum (over messages) delay and error probability as in this paper. Nevertheless, the same results hold in the latter case as discussed briefly later at the end of Section V.

IV. MAIN RESULTS

This section is divided into two parts. In Section IV-A, we provide general upper and lower bounds on capacity, and derive several of its properties. In Section IV-B, we investigate the performance limits of training-based schemes and establish their suboptimality in a certain communication regime. Since both sections can be read independently, the practically inclined reader may read Section IV-B first.

All of our results assume a uniform distribution on ν . Nevertheless, this assumption is not critical in our proofs. The results can be extended to non-uniform distributions by following the same arguments as those used to establish

⁵Note that $\mathbb{E}_{m,t}(\tau_n - t)^+$ should be interpreted as $\mathbb{E}_{m,t}((\tau_n - t)^+)$.

⁶This claim appeared in [3, p. 4515].

⁷This follows from the converse of [9, Theorem], which says that above α_o , even the codeword of a single codeword codebook is mislocated with probability tending to one.

asynchronous capacity per unit cost for non-uniform ν [5, Theorem 5].

A. General Bounds on Asynchronous Capacity

To communicate reliably, whether synchronously or asynchronously, the input-output mutual information induced by the codebook should at least be equal to the desired communication rate.

When communication is asynchronous, a decoder should, in addition, be able to discriminate between hypothesis “noise” and hypothesis “message.” These hypothesis correspond to the situations when the transmitter is idle and when it transmits a codeword, respectively. Intuitively, the more these hypotheses are statistically far apart—by means of an appropriate codebook design—the larger the level of asynchronism which can be accommodated for a given communication rate.

More specifically, a code should serve the dual purpose of minimizing the “false-alarm” and “miss” error probabilities.

Since the decoder doesn’t know ν , the decoder may output a message before even a message is sent. This is the false-alarm event and it contributes to increase the error probability—conditioned on a false-alarm the error probability is essentially one. However, false-alarms also contribute to increase the rate since it is defined with respect to the receiver’s decoding delay $\mathbb{E}(\tau - \nu)^+$. As an extreme case, by immediately decoding, *i.e.*, by setting $\tau = 1$, we get an infinite rate and error probability (asymptotically) equal to one. As it turns out, the false-alarm probability should be exponentially small to allow reliable communication under exponential asynchronism.

The miss event refers to the scenario where the decoder fails to recognize the sent message during transmission, *i.e.*, the message output looks like it was generated by noise. This event impacts the rate and, to a smaller extent, also the error probability. In fact, when the sent message is missed, the reaction delay is usually huge, of the order of A . Therefore, to guarantee a positive rate under exponential asynchronism the miss error probability should also be exponentially small.

Theorem 2 below provides a lower bound on the asynchronous capacity. The proof of this theorem is obtained by analyzing a coding scheme which performs synchronization and information transmission jointly. The codebook is a standard i.i.d. random code across time and messages and its performance is governed by the Chernoff error exponents for discriminating hypothesis “noise” from hypothesis “message.”

Theorem 2 (Lower Bound on Asynchronous Capacity):

Let $\alpha \geq 0$ and let $P \in \mathcal{P}^{\mathcal{X}}$ be some input distribution such that at least one of the following inequalities

$$\begin{aligned} D(V\|(PQ)_y) &\geq \alpha \\ D(V\|Q_*) &\geq \alpha \end{aligned}$$

holds for all distributions $V \in \mathcal{P}^{\mathcal{Y}}$, *i.e.*,

$$\min_{V \in \mathcal{P}^{\mathcal{Y}}} \max\{D(V\|(PQ)_y), D(V\|Q_*)\} \geq \alpha.$$

Then, the rate-exponent pair $(R = I(PQ), \alpha)$ is achievable. Thus, maximizing over all possible input distributions, we have the following lower bound on $\alpha(R)$ in Definition 2:

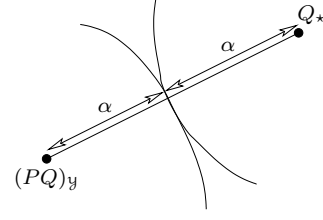


Fig. 3. If α is at most the “half-distance” between distributions $(PQ)_y$ and Q_* , then (α, R) with $R = I(PQ)$ is achievable.

$$\alpha(R) \geq \alpha_-(R) \quad R \in (0, C] \quad (5)$$

where

$$\alpha_-(R) \triangleq \max_{\substack{P \in \mathcal{P}^{\mathcal{X}} \\ I(PQ) \geq R}} \min_{V \in \mathcal{P}^{\mathcal{Y}}} \max\{D(V\|(PQ)_y), D(V\|Q_*)\}. \quad (6)$$

Theorem 2 provides a simple explicit lower bound on capacity. The distribution $(PQ)_y$ corresponds to the channel output when the input to the channel is distributed according to P . The asynchronism exponent that can be accommodated for given P and Q_* can be interpreted as being the “equidistant point” between distributions $(PQ)_y$ and Q_* , as depicted in Fig. 3. Maximizing over P such that $I(PQ) \geq R$ gives the largest such exponent that can be achieved for rate R communication.

Note that (6) is much simpler to evaluate than the lower bound given by [3, Theorem 2]. Moreover, the former is usually a better bound than the latter and it exhibits an interesting feature of $\alpha(R)$ in the high rate regime. This feature is illustrated in Example 1 to come.

Theorem 2 extends to the following continuous alphabet Gaussian setting:

Corollary 1 (Asynchronous Gaussian channel): Suppose that for a real input x the decoder receives $Y = x + Z$, where $Z \sim \mathcal{N}(0, 1)$. When there is no input to the channel, $Y = Z$, so $Q_* = \mathcal{N}(0, 1)$. The input is power constrained so that all codewords $c^n(m)$ must satisfy $\frac{1}{n} \sum_{i=1}^n c_i(m)^2 \leq p$ for a given constant $p > 0$. For this channel we have

$$\alpha(R) \geq \max_{\substack{P: I(PQ) \geq R \\ \mathbb{E}_P X^2 \leq p}} \min_V \max\{D(V\|(PQ)_y), D(V\|Q_*)\}, \quad (7)$$

for $R \in (0, C]$ where P and V in the optimization are distributions over the reals.

If we restrict the outer maximization in (7) to be over Gaussian distributions only, it can be shown that the best input has a mean μ that is as large as possible, given the rate and power constraints. More precisely, μ and R satisfy

$$R = \frac{1}{2} \ln(1 + p - \mu^2),$$

and the variance of the optimal Gaussian input is $p - \mu^2$. The intuition for choosing such parameters is that a large mean helps the decoder to distinguish the codeword from noise—since the latter has a mean equal to zero. What limits the

mean is both the power constraint and the variance needed to ensure sufficient mutual information to support communication at rate R .

Proof of Corollary 1: The proof uses a standard quantization argument similar to that in [10], and therefore we provide only a sketch of the proof. From the given the continuous time Gaussian channel, we can form a discrete alphabet channel for which we can apply Theorem 2.

More specifically, for a given constant $L > 0$, the input and the output of the channel are discretized within $[-L/2, L/2]$ into constant size Δ contiguous intervals $\Delta_i = [l_i, l_i + \Delta)$. L and Δ are chosen so that $L \rightarrow \infty$ as $\Delta \rightarrow 0$. To a given input x of the Gaussian channel is associated the quantized value $\tilde{x} = l_i + \Delta/2$ where i denotes the index of the interval Δ_i which contains x . If $x < -L/2$ or $x \geq L/2$, then \tilde{x} is defined as $-L/2$ or $L/2$, respectively. The same quantization is applied to the output of the Gaussian channel.

For each quantized channel we apply Theorem 2, then let $\Delta \rightarrow 0$ (hence $L \rightarrow \infty$). One can then verify that the achieved bound corresponds to (7), which shows that Theorem 2 also holds for the continuous alphabet Gaussian setting of Theorem 1. ■

The next result provides an upper bound to the asynchronous capacity for channels with finite synchronization threshold—see Theorem 1:

Theorem 3 (Upper Bound on Asynchronous Capacity):

For any channel Q such that $\alpha_o < \infty$, and any $R > 0$, we have that

$$\alpha(R) \leq \max_{\mathcal{S}} \min\{\alpha_1, \alpha_2\} \triangleq \alpha_+(R), \quad (8)$$

where

$$\alpha_1 \triangleq \delta(I(P_1 Q) - R + D((P_1 Q)_y \| Q_*)) \quad (9)$$

$$\alpha_2 \triangleq \min_{W \in \mathcal{P}^y | x} \max\{D(W \| Q | P_2), D(W \| Q_* | P_2)\} \quad (10)$$

with

$$\mathcal{S} \triangleq \left\{ (P_1, P_2, P'_1, \delta) \in (\mathcal{P}^x)^3 \times [0, 1] : \right. \\ \left. I(P_1 Q) \geq R, P_2 = \delta P_1 + (1 - \delta) P'_1 \right\}. \quad (11)$$

If $\alpha_o = \infty$, then

$$\alpha(R) \leq \max_{P_2} \alpha_2 \quad (12)$$

for $R \in (0, C]$.

The terms α_1 and α_2 in (8) reflect the false-alarm and miss constraints alluded to above (see discussion before Theorem 2). If $\alpha > \alpha_1$, then with high probability the noise will mimic a message before transmission starts. Instead, if $\alpha > \alpha_2$ then reliable communication at a positive rate is impossible since no code can guarantee a sufficiently low probability of missing the sent codeword.

The parameter δ in (9) and (11) essentially represents the ratio between the reaction delay $\mathbb{E}(\tau - \nu)^+$ and the blocklength—which need not coincide. Loosely speaking, for a given asynchronism level a smaller δ , or, equivalently, a smaller $\mathbb{E}(\tau - \nu)^+$, increases the communication rate at the expense of a higher false-alarm error probability. The intuition

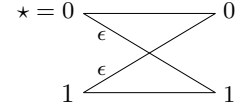


Fig. 4. A channel for which $\alpha(R)$ is discontinuous at $R = C$.

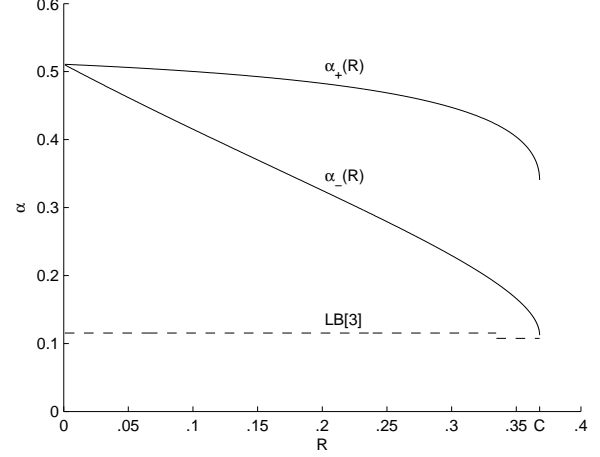


Fig. 5. Capacity upper and lower bounds on the asynchronous capacity of the channel of Fig. 4 with $\epsilon = 0.1$ and $\star = 0$. $\alpha_-(R)$ represents the lower bound given by Theorem 2, LB[3] represents the lower bound obtained in [3, Theorem 1], and $\alpha_+(R)$ represents the upper bound given by Theorem 3.

for this is that a decoder that achieves a smaller reaction delay sees, on average, “fewer” channel outputs before stopping. As a consequence, the noise is more likely to lead such a decoder into confusion. A similar tension arises between communication rate and the miss error probability. The optimization over the set \mathcal{S} attempts to strike the optimal tradeoff between the communication rate, the false-alarm and miss error probabilities, as well as the reaction delay as a fraction of the codeword length.

For channels with infinite synchronization threshold, Theorem 4 to come establishes that the bound given by (12) is actually tight.

The following examples provide some useful insights.

Example 1: Consider the binary symmetric channel depicted in Fig. 4, which has the property that when no input is supplied to the channel, the output distribution is asymmetric. For this channel, in Fig. 5 we plot the lower bound on $\alpha(R)$ given by (6) (curve $\alpha_-(R)$) and the lower bound given by [3, Theorem 1] (the dashed line LB[3]).⁸ The $\alpha_+(R)$ curve correspond to the upper bound on $\alpha(R)$ given by Theorem 3. For these plots, the channel parameter is $\epsilon = 0.1$.

The discontinuity of $\alpha(R)$ at $R = C$ (since $\alpha(R)$ is clearly equal to zero for $R > C$) implies that we do not need to back off from the synchronous capacity in order to operate under

⁸Due to the complexity of evaluating the lower bound given by [3, Theorem 1], the curves labeled LB[3] are actually upper bounds on this lower bound. We believe these bounds are fairly tight, but in any case we see that the resulting upper bounds are below the lower bounds given by (6).

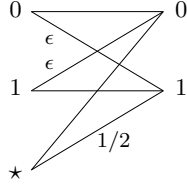


Fig. 6. Channel for which $\alpha(R)$ is continuous at $R = C$.

exponential asynchronism.⁹

Note next that the $\alpha_-(R)$ is better than LB[3] for all rates. In fact, empirical evidence suggests that $\alpha_-(R)$ is better than LB[3] in general. Additionally, note that $\alpha_-(R)$ and $\alpha_+(R)$ are not tight.

Next, we show how another binary symmetric channel has some rather different properties.

Example 2: Consider the binary symmetric channel depicted in Fig. 6, which has the property that when no input is provided to the channel the output distribution is symmetric. When used synchronously, this channel and that of Example 1 are completely equivalent, regardless of the crossover probability ϵ . Indeed, since the \star input symbol in Fig. 6 produces 0 and 1 equiprobably, this input can be ignored for coding purposes and any code for this channel achieves the same performance on the channel in Fig. 4.

However, this equivalence no longer holds when the channels are used asynchronously. To see this, we plot the corresponding upper and lower bounds on performance for this channel in Fig. 7. Comparing curve $\alpha_-(R)$ in Fig. 5 with curve $\alpha_+(R)$ in Fig. 7, we see that asynchronous capacity for the channel of Fig. 4 is always larger than that of the current example. Moreover, since there is no discontinuity in exponent at $R = C$ in our current example, the difference is pronounced at $R = C = 0.368\dots$; for the channel of Fig. 4 we have $\alpha(C) \approx 0.12 > 0$.

The discontinuity of $\alpha(R)$ at $R = C$ observed in Example 1 is in fact typical, holding in all but one special case.

Corollary 2 (Discontinuity of $\alpha(R)$ at $R = C$): We have $\alpha(C) = 0$ if and only if Q_\star corresponds to the (unique) capacity-achieving output distribution of the synchronous channel.

By Corollary 2, for the binary symmetric channel of Example 1, $\alpha(R)$ is discontinuous at $R = C$ whenever $\epsilon \neq 1/2$. To see this, note that the capacity achieving output distribution of the synchronous channel assigns equal weights to \star and 1, differently than Q_\star .

The justification for the discontinuity in Example 1 is as follows. Since the capacity-achieving output distribution of the synchronous channel (Bernoulli(1/2)) is biased with

⁹To have a better sense of what it means to be able to decode under exponential asynchronism and, more specifically, at $R = C$, consider the following numerical example. Consider a codeword length n equal to 150. Then $\alpha = .12$ yields asynchronism level $A = e^{n\alpha} \approx 6.5 \times 10^7$. If the codeword is, say, 30 centimeters long, then this means that the decoder can reliably sequentially decode the sent message, with minimal delay (were the decoder cognizant of ν , it couldn't achieve a smaller decoding delay since we operate at the synchronous capacity), within 130 kilometers of mostly noisy data!

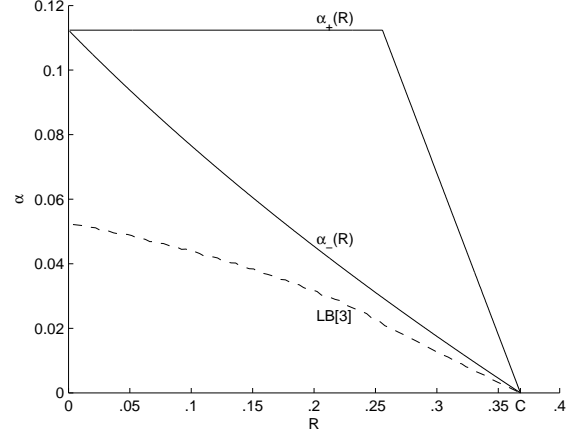


Fig. 7. Capacity upper and lower bounds on the asynchronous capacity of the channel of Fig. 6 with $\epsilon = 0.1$. $\alpha_-(R)$ represents the lower bound given by Theorem 2, LB[3] represents the lower bound obtained in [3, Theorem 1], and $\alpha_+(R)$ represents the upper bound given by Theorem 3.

respect to the noise distribution Q_\star , hypothesis “message” and “noise” can be discriminated with exponentially small error probabilities. This, in turn, enables reliable detection of the sent message under exponential asynchronism. By contrast, for the channel of Example 2, this bias no longer exists and $\alpha(R = C) = 0$. For this channel, to accommodate a positive asynchronism exponent we need to backoff from the synchronous capacity C so that the codebook output distribution can be differentiated from the noise.

Proof of Corollary 2: From Theorem 2, a strictly positive asynchronism exponent can be achieved at $R = C$ if Q_\star differs from the synchronous capacity-achieving output distribution—(6) is strictly positive for $R = C$ whenever Q_\star differs from the synchronous capacity-achieving output distribution since the divergence between two distributions is zero only if they are equal.

Conversely, suppose Q_\star is equal to the capacity-achieving output distribution of the synchronous channel. We show that for any (R, α) coding scheme where $R = C$, α is necessarily equal to zero.

From Theorem 3,

$$\alpha(R) \leq \max_{\mathcal{S}} \alpha_1$$

where \mathcal{S} and α_1 are given by (11) and (9), respectively. Since $R = C$, $I(P_1 Q) = C$, and since $Q_\star = (P_1 Q)_y$, we have $D((P_1 Q)_y || Q_\star) = 0$. Therefore, $\alpha_1 = 0$ for any δ , and we conclude that $\alpha(C) = 0$. ■

In addition to the discontinuity at $R = C$, $\alpha(R)$ may also be discontinuous at rate zero:

Corollary 3 (Discontinuity of $\alpha(R)$ at $R = 0$): If

$$\alpha_o > \max_{x \in \mathcal{X}} D(Q_\star || Q(\cdot|x)), \quad (13)$$

then $\alpha(R)$ is discontinuous at rate $R = 0$.

Example 3: Channels that satisfy (13) include those for which the following two conditions hold: \star can't produce all channel outputs, and if a channel output can be produced by \star , then it can also be produced by any other input symbol. For

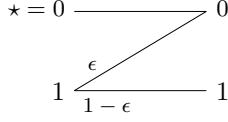


Fig. 8. Channel for which $\alpha(R)$ is discontinuous at $R = 0$, assuming $\epsilon \in (0, 1)$.

these channels (13) holds trivially; the right-hand side term is finite and the left-hand side term is infinite. The simplest such channel is the Z-channel depicted in Fig. 8 with $\epsilon \in (0, 1)$.

Note that if $\epsilon = 0$, (13) doesn't hold since both the left-hand side term and the right-hand side term are infinite. In fact, if $\epsilon = 0$ then asynchronism doesn't impact communication; rates up to the synchronous capacity can be achieved regardless of the level of asynchronism, i.e.,

$$\alpha(R) = \alpha_o = \infty \quad R \in [0, C].$$

To see this, note that by prepending a 1 to each codeword suffices to guarantee perfect synchronization without impacting rate (asymptotically).

More generally, asynchronous capacity for channels with infinite synchronization threshold is established in Theorem 4 to come.

An intuitive justification for the possible discontinuity of $\alpha(R)$ at $R = 0$ is as follows. Consider a channel where \star cannot produce all channel outputs (such as that depicted in Fig. 8). A natural encoding strategy is to start codewords with a common preamble whose possible channel outputs differ from the set of symbols that can be generated by \star . The remaining parts of the codewords are chosen to form, for instance, a good code for the synchronous channel. Whenever the decoder observes symbols that cannot be produced by noise (a clear sign of the preamble's presence), it stops and decodes the upcoming symbols. For this strategy, the probability of decoding before the message is actually sent is clearly zero. Also, the probability of wrong message isolation conditioned on correct preamble location can be made negligible by taking codewords long enough. Similarly, the probability of missing the preamble can be made negligible by using a long enough preamble. Thus, the error probability of this training-based scheme can be made negligible, regardless of the asynchronism level.

The problem arises when we add a positive rate constraint, which translates into a delay constraint. Conditioned on missing the preamble, it can be shown that the delay $(\tau - \nu)^+$ is large, in fact of order A . It can be shown that if (13) holds, the probability of missing the preamble is larger than $1/A$. Therefore, a positive rate puts a limit on the maximum asynchronism level for which reliable communication can be guaranteed, and this limit can be smaller than α_o .

We note that it is an open question whether or not $\alpha(R)$ may be discontinuous at $R = 0$ for channels that do not satisfy (13).

Theorem 4 provides an exact characterization of capacity for the class of channels with infinite synchronization threshold, i.e., whose noise distribution Q_\star cannot produce all possible channel outputs.

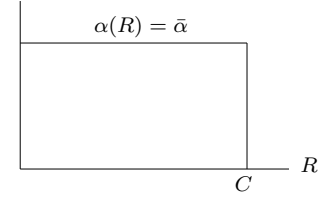


Fig. 9. Typical shape of the capacity of an asynchronous channel Q for which $\alpha_o = \infty$.

Theorem 4 (Capacity when $\alpha_o = \infty$): If $\alpha_o = \infty$, then

$$\alpha(R) = \bar{\alpha} \quad (14)$$

for $R \in (0, C]$, where

$$\bar{\alpha} \triangleq \max_{P \in \mathcal{P}^{\mathcal{X}}} \min_{W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}} \max\{D(W\|Q|P), D(W\|Q_\star|P)\}.$$

Therefore, when $\alpha_o = \infty$, $\alpha(R)$ is actually a constant that does not depend on the rate, as Fig. 9 depicts. Phrased differently, $R(\alpha) = C$ up to $\alpha = \bar{\alpha}$. For $\alpha > \bar{\alpha}$ we have $R(\alpha) = 0$.

Note that when $\alpha_o = \infty$, $\alpha(R)$ can be discontinuous at $R = 0$ since the right-hand side of (14) is upper bounded by

$$\max_{x \in \mathcal{X}} D(Q_\star \| Q(\cdot|x)),$$

which can be finite.¹⁰

We conclude this section with a result of independent interest related to synchronous communication, and which is obtained as a byproduct of the analysis used to prove Theorem 3. This result essentially says that any nontrivial fixed length codebook, i.e., that achieves a nontrivial error probability, contains a very good large (constant composition) sub-codebook, in the sense that its rate is almost the same as the original code, but its error probability decays exponentially with a suitable decoder. In the following theorem (\mathcal{C}_n, ϕ_n) denotes a standard code for a synchronous channel Q , with fixed length n codewords and decoding happening at time n .

Theorem 5: Fix a channel $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, let $q > 0$, and let $\epsilon, \gamma > 0$ be such that $\epsilon + \gamma \in (0, l)$ with $l \in (0, 1)$. If (\mathcal{C}_n, ϕ_n) is a code that achieves an error probability ϵ , then there exists an $n_o(l, \gamma, q, |\mathcal{X}|, |\mathcal{Y}|)$ such that for all $n \geq n_o$ there exists $(\mathcal{C}'_n, \phi'_n)$ such that¹¹

- 1) $\mathcal{C}'_n \subset \mathcal{C}_n$, \mathcal{C}'_n is constant composition;
- 2) the maximum error probability is less than ϵ_n where

$$\epsilon_n = 2(n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-nq^2/(2 \ln 2));$$

- 3) $\frac{\ln |\mathcal{C}'_n|}{n} \geq \frac{\ln |\mathcal{C}_n|}{n} - \gamma$.

Theorem 5 is a stronger version of [7, Corollary 1.9, p. 107] and its proof amounts to a tightening of some of the arguments in the proof of the latter, but otherwise follows it closely.

¹⁰To see this choose $W = Q_\star$ in the minimization (14).

¹¹We use $n_o(q)$ to denote some threshold index which could be explicitly given as a function of q .

B. Training-Based Schemes

Practical solutions to asynchronous communication usually separate synchronization from information transmission. We investigate a very general class of such “training-based schemes” in which codewords are composed of two parts: a preamble that is common to all codewords, followed by information symbols. The decoder first attempts to detect the preamble, then decodes the information symbols. The results in this section show that such schemes are suboptimal at least in certain communication regimes. This leads to the conclusion that the separation of synchronization and information transmission is in general not optimal.

We start by defining a general class of training-based schemes:

Definition 3 (Training-Based Scheme): A coding scheme $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ is said to be *training-based* if for some $\eta \in [0, 1]$ and all n large enough

- 1) there is a common preamble across codewords of size ηn ;
- 2) the decoding time τ_n is such that the event

$$\{\tau_n = t\},$$

conditioned on the ηn observations $Y_{t-n+1}^{t-n+\eta n}$, is independent of all other observations (i.e., Y_1^{t-n} and $Y_{t-n+\eta n+1}^{A+n-1}$).

Note that Definition 3 is in fact very general. The only restrictions are that the codewords all start with the same training sequence, and that the decoder’s decision to stop at any particular time should be based on the processing of (at most) ηn past output symbols corresponding to the length of the preamble.

In the sequel we use $\alpha^T(R)$ to denote the asynchronous capacity restricted to training based schemes.

Theorem 6 (Training-based scheme capacity bounds): Capacity restricted to training based schemes satisfies

$$\alpha_-^T(R) \leq \alpha^T(R) \leq \alpha_+^T(R) \quad R \in (0, C] \quad (15)$$

where

$$\begin{aligned} \alpha_-^T(R) &\triangleq m_1 \left(1 - \frac{R}{C}\right) \\ \alpha_+^T(R) &\triangleq \min \left\{ m_2 \left(1 - \frac{R}{C}\right), \alpha_+(R) \right\}, \end{aligned}$$

where the constants m_1 and m_2 are defined as

$$\begin{aligned} m_1 &\triangleq \max_{P \in \mathcal{P}^X} \min_{W \in \mathcal{P}^{Y|X}} \max\{D(W||Q|P), D(W||Q_*|P)\} \\ m_2 &\triangleq -\ln(\min_{y \in \mathcal{Y}} Q_*(y)), \end{aligned}$$

and where $\alpha_+(R)$ is defined in Theorem 3.

Moreover, a rate $R \in [0, C]$ training-based scheme allocates at most a fraction

$$\eta = \left(1 - \frac{R}{C}\right)$$

to the preamble.

Since $m_2 < \infty$ if and only $\alpha_o < \infty$, the upper-bound in (15) implies:

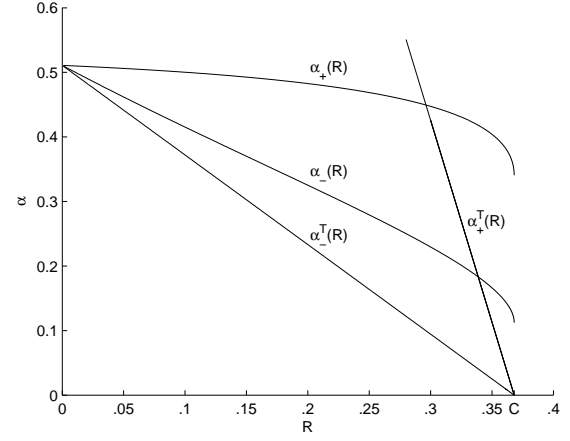


Fig. 10. Upper and lower bounds to capacity restricted to training-based schemes ($\alpha_+^T(R)$ and $\alpha_-^T(R)$, respectively) for the binary symmetric channel depicted in Fig. 4 with $\epsilon = 0.1$. $\alpha_+(R)$ and $\alpha_-(R)$ represent the capacity general upper and lower bounds given by Theorems 2 and 3.

Corollary 4 (Asynchronism in the high rate regime): For training-based schemes

$$\alpha^T(R) \xrightarrow{R \rightarrow C} 0$$

whenever $\alpha_o < \infty$.

In general, $\alpha(C) > 0$ as we saw in Corollary 2. Hence a direct consequence of Corollaries 2 and 4 is that training-based schemes are suboptimal in the high rate regime. Specifically, we have the following result.

Corollary 5 (Suboptimality of training-based schemes): There exists a channel-dependent threshold R_* such that for all $R > R_*$,

$$\alpha^T(R) < \alpha(R)$$

except possibly when Q_* corresponds to the capacity-achieving output distribution of the synchronous channel, or when the channel is degenerate, i.e., when $\alpha_o = \infty$.

The last claim of Theorem 6 says that the size of the preamble decreases (linearly) as the rate increases. This, in turn, implies that $\alpha^T(R)$ tends to zero as R approaches C . Hence, in the high rate regime most of the symbols should carry information, and the decoder should try to detect these symbols as part of the decoding process. In other words, synchronization and information transmission should be jointly performed; transmitted bits should carry information while also helping the decoder to locate the sent codeword.

If we are willing to reduce the rate, are training-based schemes still suboptimal? We do not have a definite answer to this question, but the following examples provide some insights.

Example 4: Consider the channel depicted in Fig. 4 with $\epsilon = 0.1$. In Fig. 10, we plot the upper and lower bounds to capacity restricted to training-based schemes given by Theorem 6. $\alpha_-(R)$ and $\alpha_+(R)$ represent the general lower and upper bounds to capacity given by Theorems 2 and 3; see Fig. 5.

By comparing $\alpha_-(R)$ with $\alpha_+^T(R)$ in Fig. 10 we observe that for rates above roughly 92% of the synchronous capacity C , training-based schemes are suboptimal.

For this channel, we observe that $\alpha_-(R)$ is always above $\alpha_-^T(R)$. This feature does not generalize to arbitrary crossover probabilities ϵ . Indeed, consider the channel in Fig. 4, but with an arbitrary crossover probability ϵ , and let r be an arbitrary constant such that $0 < r < 1$. From Theorem 6, training-based schemes can achieve rate asynchronism pairs (R, α) that satisfy

$$\alpha \geq m_1(1 - R/C(\epsilon)) \quad R \in (0, C(\epsilon)].$$

For the channel at hand

$$m_1 = D_B(1/2||\epsilon),$$

hence α tends to infinity as $\epsilon \rightarrow 0$, for any fixed $R \in (0, r)$ —note that $C(\epsilon) \rightarrow 1$ as $\epsilon \rightarrow 0$.

Now, consider the random coding scheme that yields Theorem 2. This scheme, which performs synchronization and information transmission jointly, achieves for any given rate $R \in [0, C]$ asynchronism exponent¹²

$$\alpha = \max_{\{P \in \mathcal{P}^x : I(PQ) \geq R\}} \min_{V \in \mathcal{P}^y} \max\{D(V||PQ)_y), D(V||Q_*)\}.$$

This expression is upper-bounded by¹³

$$\max_{P \in \mathcal{P}^x : I(PQ) \geq R} D(Q_*||PQ)_y), \quad (16)$$

which is bounded in the limit $\epsilon \rightarrow 0$ as long as $R > 0$.¹⁴ Therefore the joint synchronization-information transmission code yielding Theorem 2 can be outperformed by training-based schemes at moderate to low rate, even when the output distribution when no input is supplied is asymmetric. This shows that the general lower bound given by Theorem 2 is loose in general.

Example 5: For the channel depicted in Fig. 6 with $\epsilon = 0.1$, in Fig. 11 we plot the upper and lower bounds on capacity restricted to training-based schemes, as given by Theorem 6. For this channel it turns out that the training-based scheme upper bound $m_2(1 - R/C)$ (see Theorem 6) is loose and hence $\alpha_+^T(R) = \alpha_+(R)$ for all rates. By contrast with the example of Fig. 10, here the general lower bound $\alpha_-(R)$ is below the lower bound for the best training best schemes ($\alpha_-^T(R)$ line).

Finally, observe that, at all rates, $\alpha_+(R)$ in Fig. 11 is below $\alpha_-(R)$ (and even $\alpha_-^T(R)$) in Fig. 10. In other words, under asymmetric noise, it is possible to accommodate a much larger level of asynchronism than under symmetric noise, at all rates.

V. ANALYSIS

In this section, we establish the theorems of Section IV.

¹²The analysis of the coding scheme that yields Theorem 2 is actually tight in the sense that the coding scheme achieves (6) with equality (see proof of Theorem 2 and remark p. 14.)

¹³To see this, choose $V = Q_*$ in the minimization.

¹⁴Let $P^* = P^*(Q)$ be an input distribution P that maximizes (16) for a given channel. Since $R \leq I(P^*Q) \leq H(P^*)$, P^* is uniformly bounded away from 0 and 1 for all $\epsilon \geq 0$. This implies that (16) is bounded in the limit $\epsilon \rightarrow 0$.

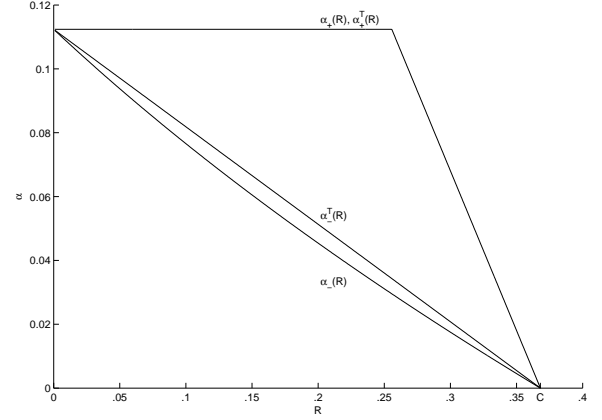


Fig. 11. Lower bound ($\alpha_-^T(R)$) to capacity restricted to training-based schemes for the channel of Fig. 6. $\alpha_+(R)$ and $\alpha_-(R)$ represent the capacity general upper and lower bounds given by Theorems 2 and 3. For this channel the training upper bound ($\alpha_+^T(R)$) coincides with $\alpha_+(R)$, and hence is not plotted separately.

A. Proof of Theorem 2

Let $\alpha \geq 0$ and $P \in \mathcal{P}^x$ satisfy the assumption of the theorem, i.e., be such that at least one of the following inequalities holds

$$\begin{aligned} D(V||PQ)_y) &\geq \alpha \\ D(V||Q_*) &\geq \alpha \end{aligned} \quad (17)$$

for all distributions $V \in \mathcal{P}^y$, and let $A_n = e^{n(\alpha - \epsilon)}$.

The proof is based on a random coding argument associated with the following communication strategy. The codebook $\mathcal{C} = \{c^n(m)\}_{m=1}^M$ is randomly generated so that all $c_i(m)$, $i \in \{1, 2, \dots, n\}$, $m \in \{1, 2, \dots, M\}$, are i.i.d. according to P . The sequential decoder operates according to a two-step procedure. The first step consists in making an coarse estimate of the location of the sent codeword. Specifically, at time t the decoder tries to determine whether the last n output symbols are generated by noise or by some codeword on the basis of their empirical distribution $\hat{P} = \hat{P}_{y_{t-n+1}^t}$. If $D(\hat{P}||Q_*) < \alpha$, \hat{P} is declared a “noise type,” the decoder moves to time $t+1$, and repeats the procedure, i.e., tests whether $\hat{P}_{y_{t-n+2}^{t+1}}$ is a noise type. If, instead, $D(\hat{P}||Q_*) \geq \alpha$, the decoder marks the current time as the beginning of the “decoding window,” and proceeds to the second step of the decoding procedure.

The second step consists in exactly locating and identifying the sent codeword. Once the beginning of the decoding window has been marked, the decoder makes a decision the first time that the previous n symbols are jointly typical with one of the codewords. If no such time is found within n successive time steps, the decoder stops and declares a random message. The typicality decoder operates as follows.¹⁵ Let P_m be the probability measure induced by codeword $c^n(m)$ and

¹⁵In the literature this decoder is often referred to as the “strong typicality” decoder.

the channel, i.e.,

$$P_m(a, b) \triangleq \hat{P}_{c^n(m)}(a)Q(b|a) \quad (a, b) \in \mathcal{X} \times \mathcal{Y}. \quad (18)$$

At time t , the decoder computes the empirical distributions \hat{P}_m induced by $c^n(m)$ and the n output symbols y_{t-n+1}^t for all $m \in \{1, 2, \dots, M\}$. If

$$|\hat{P}_{c^n(m), y_{t-n+1}^t}(a, b) - P_m(a, b)| \leq \mu$$

for all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ and a unique index m , the decoder declares message m as the sent message. Otherwise, it moves one step ahead and repeats the second step of the decoding procedure on the basis of y_{t-n+2}^{t+1} , i.e., it tests whether y_{t-n+2}^{t+1} is typical with a codeword.

At the end of the asynchronism time window, i.e., at time $A_n + n - 1$, if $\hat{P}_{A_n + n - 1}^{A_n + n - 1}$ is either a noisy type or if it is typical with none of the codewords, the decoder declares a message at random.

Throughout the argument we assume that the typicality parameter μ is a negligible, strictly positive quantity.

We first show that, on average, a randomly chosen codebook combined with the sequential decoding procedure described above achieves the rate-exponent pairs (R, α) claimed by the theorem. This, as we show at the end of the proof, implies the existence of a nonrandom codebook that, together with the above decoding procedure, achieves any pair (R, α) claimed by the theorem.

Let $\ln M/n = I(PQ) - \epsilon$, $\epsilon > 0$. We first compute the average, over messages and codes, expected reaction delay and probability of error. These quantities, by symmetry of the encoding and decoding procedures, are the same as the average over codes expected reaction delay and probability of error conditioned on the sending of a particular message. Below, expected reaction delay and error probability are computed conditioned on the sending of message $m = 1$.

Define the following events:

$$\mathcal{E}_1 = \{D(\hat{P}_{Y_{\nu+n-1}^{\nu+n-1}} \| Q_*) < \alpha, \text{ i.e., } \hat{P}_{Y_{\nu+n-1}^{\nu+n-1}} \text{ is a "noise type"}\},$$

$$\mathcal{E}_2 = \{Y_{\nu+n-1}^{\nu+n-1} \text{ is not typical with } C^m(1)\},$$

$$\mathcal{E}_3 = \{D(\hat{P}_{Y_{t-n+1}^t} \| Q_*) \geq \alpha \text{ for some } t < \nu\}.$$

For the reaction delay we have

$$\begin{aligned} \mathbb{E}_1(\tau_n - \nu)^+ &= \mathbb{E}_1[(\tau_n - \nu)^+ \mathbb{1}(\tau_n \geq \nu + 2n)] \\ &\quad + \mathbb{E}_1[(\tau_n - \nu)^+ \mathbb{1}(\nu + n \leq \tau_n < \nu + 2n)] \\ &\quad + \mathbb{E}_1[(\tau_n - \nu)^+ \mathbb{1}(\tau_n < \nu + n)] \\ &\leq (A_n + n - 1) \mathbb{P}_1(\tau_n \geq \nu + 2n) \\ &\quad + 2n \mathbb{P}_1(\nu + n \leq \tau_n < \nu + 2n) + n, \end{aligned} \quad (19)$$

where the subscript 1 in \mathbb{E}_1 and \mathbb{P}_1 indicates conditioning on the event that message $m = 1$ is sent. The two probability terms on the right-hand side of the second inequality of (19) are bounded as follows.

The term $\mathbb{P}_1(\tau_n \geq \nu + 2n)$ is upper bounded by the probability that the decoding window starts after time $\nu + n - 1$. This, in turn, is upper bounded by the probability of the event

that, at time $\nu + n - 1$, the last n output symbols induce a noisy type. Therefore, we have

$$\begin{aligned} \mathbb{P}_1(\tau_n \geq \nu + 2n) &\leq \mathbb{P}_1(\mathcal{E}_1) \\ &\leq \sum_{\{V \in \mathcal{P}_n^{\mathcal{Y}}: D(V \| Q_*) \leq \alpha\}} e^{-nD(V \| (PQ)_y)} \\ &\leq \sum_{\{V \in \mathcal{P}_n^{\mathcal{Y}}: D(V \| Q_*) \leq \alpha\}} e^{-n\alpha} \\ &\leq \text{poly}(n) e^{-n\alpha}, \end{aligned} \quad (20)$$

where the second inequality follows from the definition of the event \mathcal{E}_1 and Fact 2; where the third inequality follows from (17) (which implies that if $D(V \| Q_*) \leq \alpha$ then necessarily $D(V \| (PQ)_y) \geq \alpha$); and where the fourth inequality follows from Fact 1.

The probability $\mathbb{P}_1(\nu + n \leq \tau_n < \nu + 2n)$ is at most the probability that the decoder has not stopped by time $\nu + n - 1$. This probability, in turn, is at most the probability that, at time $\nu + n - 1$, the last n output symbols either induce a noisy type, or are not typical with the sent codeword $C^n(1)$ (recall that message $m = 1$ is sent). By union bound we get

$$\begin{aligned} \mathbb{P}_1(\nu + n \leq \tau_n < \nu + 2n) &\leq \mathbb{P}_1(\tau_n \geq \nu + n) \\ &\leq \mathbb{P}_1(\mathcal{E}_1) + \mathbb{P}_1(\mathcal{E}_2) \\ &\leq \text{poly}(n) e^{-n\alpha} + o(1) \\ &= o(1) \quad (n \rightarrow \infty), \end{aligned} \quad (21)$$

where we used the last three computation steps of (20) to bound $\mathbb{P}_1(\mathcal{E}_1)$, and where we used [7, Lemma 2.12, p. 34] to show that $\mathbb{P}_1(\mathcal{E}_2)$ tends to zero as n tends to infinity. From (19), (20), and (21), we deduce that

$$\mathbb{E}_1(\tau_n - \nu)^+ \leq n(1 + o(1)) \quad (n \rightarrow \infty)$$

since $A_n = e^{n(\alpha - \epsilon)}$, by assumption.

We now compute $\mathbb{P}_1(\mathcal{E})$, the average error probability conditioned on sending message $m = 1$. We have

$$\begin{aligned} \mathbb{P}_1(\mathcal{E}) &= \mathbb{P}_1(\mathcal{E} \cap \{\tau_n < \nu\}) \\ &\quad + \mathbb{P}_1(\mathcal{E} \cap \{\nu \leq \tau_n \leq \nu + n - 1\}) \\ &\quad + \mathbb{P}_1(\mathcal{E} \cap \{\tau_n \geq \nu + n\}) \\ &\leq \mathbb{P}_1(\tau_n < \nu) + \mathbb{P}_1(\mathcal{E} \cap \{\nu \leq \tau_n \leq \nu + n - 1\}) \\ &\quad + \mathbb{P}_1(\tau_n \geq \nu + n) \\ &\leq \mathbb{P}_1(\mathcal{E}_3) + \mathbb{P}_1(\mathcal{E} \cap \{\nu \leq \tau_n \leq \nu + n - 1\}) \\ &\quad + o(1) \quad (n \rightarrow \infty), \end{aligned} \quad (22)$$

where for the last inequality we used the definition of \mathcal{E}_3 and upper bounded $\mathbb{P}_1(\tau \geq \nu + n)$ using the last three computation steps of (21).

For $\mathbb{P}_1(\mathcal{E}_3)$, we have

$$\begin{aligned}
\mathbb{P}_1(\mathcal{E}_3) &= \mathbb{P}(\cup_{t < \nu} \{D(\hat{P}_{Y_{t-n+1}^t} \| Q_*) \geq \alpha\}) \\
&\leq A_n \sum_{\{V \in \mathcal{P}_n^{\mathcal{X}}: D(V \| Q_*) \geq \alpha\}} e^{-nD(V \| Q_*)} \\
&\leq A_n \sum_{\{V \in \mathcal{P}_n^{\mathcal{X}}: D(V \| Q_*) \geq \alpha\}} e^{-n\alpha} \\
&\leq A_n e^{-n\alpha} \text{poly}(n) \\
&= o(1) \quad (n \rightarrow \infty)
\end{aligned} \tag{23}$$

where the first inequality in (23) follows from the union bound over time and Fact 2; where the third inequality follows from Fact 1; and where the last equality holds since $A_n = e^{n(\alpha-\epsilon)}$, by assumption.

We now show that

$$\mathbb{P}_1(\mathcal{E} \cap \{\nu \leq \tau_n \leq \nu + n - 1\}) = o(1) \quad (n \rightarrow \infty), \tag{24}$$

which, together with (22) and (23), shows that $\mathbb{P}_1(\mathcal{E})$ goes to zero as $n \rightarrow \infty$.

We have

$$\begin{aligned}
&\mathbb{P}_1(\mathcal{E} \cap \{\nu \leq \tau_n \leq \nu + n - 1\}) \\
&= \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-1} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3\}) \\
&\quad + \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-1} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}) \\
&\leq \mathbb{P}_1(\mathcal{E}_3) + \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-1} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}) \\
&\leq o(1) + \mathbb{P}_1(\{\mathcal{E} \cap \{\tau_n = \nu + n - 1\}\}) \\
&\quad + \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-2} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}) \\
&\leq o(1) + o(1) \\
&\quad + \mathbb{P}_1(\cup_{t=\nu}^{\nu+n-2} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}) \quad (n \rightarrow \infty)
\end{aligned} \tag{25}$$

where the second inequality follows from (23); where the fourth inequality follows from the definition of event \mathcal{E}_2 ; and where the third inequality follows from the fact that, given the correct codeword location, i.e., $\tau_n = \nu + n - 1$, the typicality decoder guarantees vanishing error probability since we assumed that $\ln M/n = I(PQ) - \epsilon$ (see [7, Chapter 2.1]).

The event $\{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}$, with $\nu \leq t \leq \nu + n - 2$, happens when a block of n consecutive symbols, received between $\nu - n + 1$ and $\nu + n - 2$, is jointly typical with a codeword other than the sent codeword $C^n(1)$. Consider a block Y^n in this range, and let $J \in \mathcal{P}_n^{\mathcal{X}, \mathcal{Y}}$ be a typical joint type, i.e.

$$|J(x, y) - P(x)Q(y|x)| \leq \mu$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ —recall that $\mu > 0$ is the typicality parameter, which we assume to be a negligible quantity throughout the proof.

For some $1 \leq k \leq n - 1$, the first k symbols of block Y^n are generated by noise, and the remaining $n - k$ symbols are generated by the sent codeword, i.e., corresponding to $m = 1$. Thus, Y^n is independent of any unsent codeword $C^n(m)$. The probability that $C^n(m)$, $m \neq 1$, together with Y^n yields a

particular type J is upper bounded as follows:

$$\begin{aligned}
&\mathbb{P}(\hat{P}_{C^n(m), Y^n} = J) \\
&= \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}(Y^n = y^n) \sum_{x^n: \hat{P}_{x^n, y^n} = J} \mathbb{P}(X^n = x^n) \\
&= \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}(Y^n = y^n) \sum_{x^n: \hat{P}_{x^n, y^n} = J} e^{-n(H(J_{\mathcal{X}}) + D(J_{\mathcal{X}} \| P))} \\
&\leq \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}(Y^n = y^n) e^{-nH(J_{\mathcal{X}})} |\{x^n : \hat{P}_{x^n, y^n} = J\}| \\
&\leq \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}_1(Y^n = y^n) e^{-nH(J_{\mathcal{X}})} e^{nH(J_{\mathcal{X}|Y})} \\
&\leq e^{-nI(J)},
\end{aligned} \tag{26}$$

where $H(J_{\mathcal{X}})$ denotes the entropy of the left marginal of J ,

$$H(J_{\mathcal{X}|Y}) \triangleq - \sum_{y \in \mathcal{Y}} J_Y(y) \sum_{x \in \mathcal{X}} J_{X|Y}(x|y) \ln J_{X|Y}(x|y),$$

and where $I(J)$ denotes the mutual information induced by J .

The first equality in (26) follows from the independence of $C^n(m)$ and Y^n , the second equality follows from [11, Theorem 11.1.2, p. 349], and the second inequality follows from [7, Lemma 2.5, p. 31].

It follows that the probability that an unsent codeword $C^n(m)$ together with Y^n yields a type J that is typical, i.e., close to PQ , is upper bounded as

$$\mathbb{P}_1(\hat{P}_{C^n(m), Y^n} = J) \leq e^{-n(I(PQ) - \epsilon/2)}$$

for all n large enough, by continuity of the mutual information.¹⁶

Note that the set of inequalities (26) holds for any block of n consecutive output symbols Y^n that is independent of codeword $C^n(m)$.¹⁷ Hence, from the union bound, it follows that

$$\begin{aligned}
&\mathbb{P}_1(\cup_{t=\nu}^{\nu+n-2} \{\mathcal{E} \cap \{\tau_n = t\} \cap \mathcal{E}_3^c\}) \\
&\leq n \sum_{m \neq 1} \sum_{\{J \in \mathcal{P}_{\mathcal{X}, \mathcal{Y}}^n: \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \\ &\quad |J(x, y) - P(x)Q(y|x)| \leq \mu\}} \mathbb{P}(\hat{P}_{C^n(m), Y^n} = J) \\
&\leq n M e^{-n(I(PQ) - \epsilon/2)} \text{poly}(n) \\
&\leq e^{-n\epsilon/2} \text{poly}(n),
\end{aligned} \tag{27}$$

where the second inequality follows from Fact 1, and where the third inequality follows from the assumption that $\ln M/n = I(PQ) - \epsilon$. Combining (27) with (25) yields (24).

So far, we have proved that a random codebook has a decoding delay averaged over messages that is at most $n(1 + o(1))$ ($n \rightarrow \infty$), and an error probability averaged over messages that vanishes as $n \rightarrow \infty$, whenever $A_n = e^{n(\alpha-\epsilon)}$, $\epsilon > 0$. This, as we now show, implies the existence of nonrandom codebooks achieving the same performance, yielding the desired result. The expurgation arguments we use are standard

¹⁶The typicality parameter $\mu = \mu(\epsilon) > 0$ is chosen small enough so that this inequality holds.

¹⁷Note that the fact that Y^n is partly generated by noise and partly by the sent codeword $C^n(1)$ is not used to establish (26).

and in the same spirit as those given in [11, p. 203-204] or [12, p. 151].

For a particular codebook \mathcal{C}_n , let $\mathbb{P}(\mathcal{E}|\mathcal{C}_n)$ and $\mathbb{E}((\tau_n - \nu)^+|\mathcal{C}_n)$ be the average, over messages, error probability and reaction delay, respectively. We have proved that for any $\epsilon > 0$,

$$\mathbb{E}(\mathbb{E}(\tau_n - \nu)^+|\mathcal{C}_n)) \leq n(1 + \epsilon)$$

and

$$\mathbb{E}(\mathbb{P}(\mathcal{E}|\mathcal{C}_n)) \leq \epsilon$$

for all n large enough.

Define events

$$\mathcal{A}_1 = \{\mathbb{E}(\tau_n - \nu)^+|\mathcal{C}_n) \leq n(1 + \epsilon)^2\},$$

and

$$\mathcal{A}_2 = \{\mathbb{P}(\mathcal{E}|\mathcal{C}_n) \leq \epsilon k\}$$

where k is arbitrary.

From Markov's inequality it follows that¹⁸

$$\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2) \geq 1 - \frac{1}{1 + \epsilon} - \frac{1}{k}.$$

Letting k be large enough so that the right-hand side of the above inequality is positive, we deduce that there exists a particular code \mathcal{C}_n such that

$$\mathbb{E}(\tau_n - \nu)^+|\mathcal{C}_n) \leq n(1 + \epsilon)^2$$

and

$$\mathbb{P}(\mathcal{E}|\mathcal{C}_n) \leq \epsilon k.$$

We now remove from \mathcal{C}_n codewords with poor reaction delay and error probability. Repeating the argument above with the fixed code \mathcal{C}_n , we see that a positive fraction of the codewords of \mathcal{C}_n have expected decoding delay at most $n(1 + \epsilon)^3$ and error probability at most ϵk^2 . By only keeping this set of codewords, we conclude that for any $\epsilon > 0$ and all n large enough, there exists a rate $R = I(PQ) - \epsilon$ code operating at asynchronism level $A = e^{(\alpha - \epsilon)n}$ with maximum error probability less than ϵ . ■

Remark 2: It is possible to somewhat strengthen the conclusion of Theorem 2 in two ways. First, it can be strengthened by observing that what we actually proved is that the error probability not only vanishes but does so exponentially in n .¹⁹ Second, it can be strengthened by showing that the proposed random coding scheme achieves (6) with equality. A proof is deferred to Appendix A.

B. Proof of Theorem 3

We show that any rate $R > 0$ coding scheme operates at an asynchronism α bounded from above by $\max_{\mathcal{S}} \min\{\alpha_1, \alpha_2\}$, where \mathcal{S} , α_1 , and α_2 are defined in the theorem's statement.

We prove Theorem 3 by establishing the following four claims.

The first claim says that, without loss of generality, we may restrict ourselves to constant composition codes. Specifically, it

is possible to expurgate an arbitrary code to make it of constant composition while impacting (asymptotically) neither the rate nor the asynchronism exponent the original code is operating at. In more detail, the expurgated codebook is such that all codewords have the same type, and also so that all codewords have the same type over the first Δ_n symbols (recall that $\Delta_n \triangleq \max_m \mathbb{E}(\tau_n - \nu)^+$). The parameter δ in Theorem 3 corresponds to the ratio Δ_n/n , and P_1 and P_2 correspond to the empirical types over the first Δ_n symbols and the whole codeword (all n symbols), respectively.

Fix an arbitrarily small constant $\epsilon > 0$.

Claim 1: Given any coding scheme $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ achieving (R, α) with $R > 0$ and $\alpha > 0$, there exists a second coding scheme $\{(\mathcal{C}'_n, (\tau_n, \phi_n))\}_{n \geq 1}$ achieving (R, α) that is obtained by expurgation, i.e., $\mathcal{C}'_n \subset \mathcal{C}_n$, $n = 1, 2, \dots$, and that has constant composition with respect to some distribution P_n^1 over the first

$$d(n) \triangleq \min\{\lfloor (1 + \epsilon)\Delta_n \rfloor, n\} \quad (28)$$

symbols, and constant composition with respect to some distribution P_n^2 over n symbols. (Hence, if $\lfloor (1 + \epsilon)\Delta_n \rfloor \geq n$, then $P_n^1 = P_n^2$.) Distributions P_n^1 and P_n^2 satisfy Claims 2 – 4 below.

Distribution P_n^1 plays the same role as the codeword distribution for synchronous communication. As such it should induce a large enough input-output channel mutual information to support rate R communication.

Claim 2: For all n large enough

$$R \leq I(P_n^1 Q)(1 + \epsilon).$$

Distribution P_n^2 is specific to asynchronous communication. Intuitively, P_n^2 should induce an output distribution that is sufficiently different from pure noise so that to allow a decoder to distinguish between noise and any particular transmitted message when the asynchronism level corresponds to α . Proper message detection means that the decoder should not overreact to a sent codeword (i.e., declare a message before even it is sent), but also not miss the sent codeword. As an extreme case, it is possible to achieve a reaction delay $\mathbb{E}(\tau - \nu)^+$ equal to zero by setting $\tau = 1$, at the expense of a large probability of error. In contrast, one clearly minimizes the error probability by waiting until the end of the asynchronism window, i.e., by setting $\tau = A_n + n - 1$, at the expense of the rate, which will be negligible in this case.

The ability to properly detect only a single codeword with type P_n^2 is captured by condition $\alpha \leq \alpha_2$ where α_2 is defined in the theorem's statement. This condition is equivalently stated as:

Claim 3: For any $W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ and for all n large enough, at least one of the following two inequalities holds

$$\begin{aligned} \alpha &< D(W \| Q_\star | P_n^2) + \epsilon, \\ \alpha &< D(W \| Q | P_n^2) + \epsilon. \end{aligned}$$

As it turns out, if the synchronization threshold is finite, P_n^1 plays also a role in the decoder's ability to properly detect the transmitted message. This is captured by condition $\alpha \leq \alpha_1$ where α_1 is defined in the theorem's statement. Intuitively,

¹⁸Probability here is averaged over randomly generated codewords.

¹⁹Note that the error probability of the typicality decoder given the correct message location, i.e., $\mathbb{P}(\mathcal{E} \cap \{\tau_n = \nu + n - 1\})$, is exponentially small in n [7, Chapter 2].

α_1 relates to the probability that the noise produces a string of length n that looks typical with the output of a *randomly selected codeword*. If $\alpha > \alpha_1$, the noise produces many such strings with high probability, which implies a large probability of error.

Claim 4: For all n large enough,

$$\alpha \leq \frac{d(n)}{n} (I(P_n^1 Q) - R + D((P_n^1 Q)_Y \| Q_*)) + \epsilon$$

provided that $\alpha_o < \infty$.

Note that, by contrast with the condition in Claim 3, the condition in Claim 4 depends also on the communication rate since the error yielding to the latter condition depends on the number of codewords.

Before proving the above claims, we show how they imply Theorem 3. The first part of the Theorem, i.e., when $\alpha_o < \infty$, follows from Claims 1-4. To see this, note that the bounds α_1 and α_2 in the Theorem correspond to the bounds of Claims 3 and 4, respectively, maximized over P_n^1 and P_n^2 . The maximization is subjected to the two constraints given by Claims 1 and 2: P_n^1 and P_n^2 are the empirical distributions of the codewords of \mathcal{C}'_n over the first δn symbols ($\delta \in [0, 1]$), and over the entire codeword length, respectively, and condition $R \leq I(P_n^1 Q)(1 + \epsilon)$ must be satisfied. Since $\epsilon > 0$ is arbitrary, the result then follows by taking the limit $\epsilon \downarrow 0$ on the above derived bound on α .

Similarly, the second part of Theorem 3, i.e., when $\alpha_o = \infty$, is a consequence of Claim 3 only.

We now prove the claims. As above, $\epsilon > 0$ is supposed to be an arbitrarily small constant.

Proofs of Claims 1 and 2: We show that for all n large enough, we have

$$\frac{R - \epsilon}{1 + \epsilon} \leq \frac{\ln |\mathcal{C}'_n|}{d(n)} \leq I(P_n^1 Q) + \epsilon, \quad (29)$$

where \mathcal{C}'_n is a subset of codewords from \mathcal{C}_n that have constant composition P_n^1 over the first $d(n)$ symbols, where $d(n)$ is defined in (28), and constant composition P_n^2 over n symbols. This is done via an expurgation argument in the spirit of [12, p. 151] and [11, p. 203-204].

We first show the left-hand side inequality of (29). Since $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ achieves a rate R , by definition (see Definition 1) we have

$$\frac{\ln |\mathcal{C}_n|}{\Delta_n} \geq R - \epsilon/2$$

for all n large enough. Therefore,

$$\frac{\ln |\mathcal{C}_n|}{d(n)} \geq \frac{R - \epsilon/2}{1 + \epsilon}$$

for all n large enough.

Now, group the codewords of \mathcal{C}_n into families such that elements of the same family have the same type over the first $d(n)$ symbols. Let \mathcal{C}''_n be the largest such family and let P_n^1 be its type. Within \mathcal{C}''_n , consider the largest subfamily \mathcal{C}'_n of codewords that have constant composition over n symbols, and let P_n^2 be its type (hence, all the codewords in \mathcal{C}'_n have common type P_n^1 over $d(n)$ symbols and common type P_n^2 over n symbols).

By assumption, $R > 0$, so \mathcal{C}_n has a number of codewords that is exponential in Δ_n . Due to Fact 1, to establish the left-hand side inequality of (29), i.e., to show that \mathcal{C}'_n achieves essentially the same rate as \mathcal{C}_n , it suffices to show that the number of subfamilies in \mathcal{C}'_n is bounded by a polynomial in Δ_n . We do this assuming that $\alpha_o < \infty$ and that Claim 4 (to be proved) holds.

By assumption, $\alpha_o < \infty$, and thus from Theorem 1 we have that $D((PQ)_Y \| Q_*) < \infty$ for any input distribution P . Using Claim 4 and the assumption that $\alpha > 0$, we deduce that $\liminf_{n \rightarrow \infty} d(n)/n > 0$, which implies that n cannot grow faster than linearly in Δ_n . Therefore, Fact 1 implies that the number of subfamilies of \mathcal{C}'_n is bounded by a polynomial in Δ_n .

We now prove the right-hand side inequality of (29). Letting \mathcal{E}^c denote the event of a correct decoding, Markov's inequality implies that for every message index m ,

$$\begin{aligned} \mathbb{P}_m(\{(\tau_n - \nu)^+ \leq (1 + \epsilon)\Delta_n\} \cap \mathcal{E}^c) \\ \geq 1 - \frac{\mathbb{E}_m(\tau_n - \nu)^+}{\Delta_n} \frac{1}{1 + \epsilon} - \mathbb{P}_m(\mathcal{E}) \\ \geq 1 - \frac{1}{1 + \epsilon} - \mathbb{P}_m(\mathcal{E}), \end{aligned} \quad (30)$$

since $\Delta_n \triangleq \max_m \mathbb{E}_m(\tau_n - \nu)^+$. The right-hand side of (30) is strictly greater than zero for n large enough because an (R, α) coding scheme achieves a vanishing maximum error probability as $n \rightarrow \infty$. This means that \mathcal{C}'_n is a good code for the synchronous channel, i.e., for $A = 1$. More precisely, the codebook formed by truncating each codeword in \mathcal{C}'_n to include only the first $d(n)$ symbols achieves a probability of error (asymptotically) bounded away from one with a suitable decoding function. This implies that the right-hand side of (29) holds for n large enough by [7, Corollary 1.4, p. 104]. ■

In establishing the remaining claims of the proof, unless otherwise stated, whenever we refer to a codeword it is assumed to belong to codebook \mathcal{C}'_n . Moreover, for convenience, and with only minor abuse of notation, we let M denote the number of codewords in \mathcal{C}'_n .

Proof of Claim 3: We fix $W \in \mathcal{P}^Y|X$ and show that for all n large enough, at least one of the two inequalities

$$D(W \| Q | P_n^2) > \alpha - \epsilon,$$

$$D(W \| Q_* | P_n^2) > \alpha - \epsilon,$$

must hold. To establish this, it may be helpful to interpret W as the true channel behavior during the information transmission period, i.e., as the conditional distribution induced by the transmitted codeword and the corresponding channel output. With this interpretation, $D(W \| Q | P_n^2)$ represents the large deviation exponent of the probability that the underlying channel Q behaves as W when codeword distribution is P_n^2 , and $D(W \| Q_* | P_n^2)$ represents the large deviation exponent of the probability that the noise behaves as W when codeword distribution is P_n^2 . As it turns out, if both the above inequalities are reversed for a certain W , the asynchronism exponent is too large. In fact, in this case both the transmitted message and pure noise are very likely to produce such a W . This, in turn

will confuse the decoder. It will either miss the transmitted codeword or stop before even the actual codeword is sent.

In the sequel, we often use the shorthand notation $\mathcal{T}_W(m)$ for $\mathcal{T}_W^n(c^n(m))$.

Observe first that if n is such that

$$\mathbb{P}_m(Y_\nu^{\nu+n-1} \in \mathcal{T}_W(m)) = 0, \quad (31)$$

then

$$D(W\|Q|P_n^2) = \infty,$$

by Fact 3. Similarly, observe that if n is such that

$$\mathbb{P}_*(Y_\nu^{\nu+n-1} \in \mathcal{T}_W(m)) = 0, \quad (32)$$

where \mathbb{P}_* denotes the probability under pure noise (i.e., the Y_i 's are i.i.d. according to Q_*), then

$$D(W\|Q_*|P_n^2) = \infty.$$

Since the above two observations hold regardless of m (because all codewords in \mathcal{C}'_n have the same type), Claim 3 holds trivially for any value of n for which (31) or (32) is satisfied.

In the sequel, we thus restrict our attention to values of n for which

$$\mathbb{P}_m(Y_\nu^{\nu+n-1} \in \mathcal{T}_W(m)) \neq 0 \quad (33)$$

and

$$\mathbb{P}_*(Y_\nu^{\nu+n-1} \in \mathcal{T}_W(m)) \neq 0. \quad (34)$$

Our approach is to use a change of measure to show that if Claim 3 does not hold, then the expected reaction delay grows exponentially with n , implying that the rate is asymptotically equal to zero. To see this, note that any coding scheme that achieves vanishing error probability cannot have $\ln M$ grow faster than linearly with n , simply because of the limitations imposed by the capacity of the synchronous channel. Therefore, if $\mathbb{E}(\tau_n - \nu)^+$ grows exponentially with n , the rate goes to zero exponentially with n . And note that for $\mathbb{E}(\tau_n - \nu)^+$ to grow exponentially, it suffices that $\mathbb{E}_m(\tau_n - \nu)^+$ grows exponentially for at least one message index m , since $\Delta_n = \max_m \mathbb{E}_m(\tau_n - \nu)^+$ by definition.

To simplify the exposition and avoid heavy notation, in the following arguments we disregard discrepancies due to the rounding of noninteger quantities. We may, for instance, treat A/n as an integer even if A is not a multiple of n . This has no consequences on the final results, as these discrepancies vanish when we consider code with blocklength n tending to infinity.

We start by lower bounding the reaction delay as²⁰

$$\begin{aligned} \Delta_n &\triangleq \max_m \frac{1}{A} \sum_{t=1}^A \mathbb{E}_{m,t}(\tau_n - t)^+ \\ &\geq \frac{1}{3} \sum_{t=1}^{A_n/3} \mathbb{P}_{m,t}((\tau_n - t)^+ \geq A_n/3) \\ &\geq \frac{1}{3} \sum_{t=1}^{A_n/3} \mathbb{P}_{m,t}(\tau_n \geq t + A_n/3) \end{aligned}$$

²⁰Recall that the subscripts m, t indicate conditioning on the event that message m starts being sent at time t .

$$\geq \frac{1}{3} \sum_{t=1}^{A_n/3} \mathbb{P}_{m,t}(\tau_n \geq 2A_n/3), \quad (35)$$

where for the first inequality we used Markov's inequality. The message index m on the right-hand side of (35) will be specified later; for now it may correspond to any message.

We lower bound each term $\mathbb{P}_{m,t}(\tau_n \geq 2A_n/3)$ in the above sum as

$$\begin{aligned} \mathbb{P}_{m,t}(\tau_n \geq 2A_n/3) &\geq \mathbb{P}_{m,t}(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ &\quad \times \mathbb{P}_{m,t}(Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ &\geq \mathbb{P}_{m,t}(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ &\quad \times e^{-nD_1} \text{poly}(n), \end{aligned} \quad (36)$$

where $D_1 \triangleq D(W\|Q|P_n^2)$, and where the second inequality follows from Fact 3.²¹

The key step is to apply the change of measure

$$\begin{aligned} \mathbb{P}_{m,t}(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ = \mathbb{P}_*(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)). \end{aligned} \quad (37)$$

To see that (37) holds, first note that for any y^n

$$\begin{aligned} \mathbb{P}_{m,t}(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} = y^n) \\ = \mathbb{P}_*(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} = y^n) \end{aligned}$$

since distribution $\mathbb{P}_{m,t}$ and \mathbb{P}_* differ only over channel outputs Y_t^{t+n-1} .

Next, since sequences inside $\mathcal{T}_W(m)$ are permutations of each other

$$\begin{aligned} \mathbb{P}_{m,t}(Y_t^{t+n-1} = y^n \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) &= \frac{1}{|\mathcal{T}_W(m)|} \\ &= \mathbb{P}_*(Y_t^{t+n-1} = y^n \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)), \end{aligned}$$

we get

$$\begin{aligned} \mathbb{P}_{m,t}(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ = \sum_{y^n \in \mathcal{T}_W(m)} \mathbb{P}_{m,t}(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} = y^n) \\ \quad \times \mathbb{P}_{m,t}(Y_t^{t+n-1} = y^n \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ = \sum_{y^n \in \mathcal{T}_W(m)} \mathbb{P}_*(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} = y^n) \\ \quad \times \mathbb{P}_*(Y_t^{t+n-1} = y^n \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ = \mathbb{P}_*(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)). \end{aligned}$$

This proves (37). Substituting (37) into the right-hand side of (36) and using (35), we get

$$\begin{aligned} \Delta_n &\geq e^{-nD_1} \text{poly}(n) \\ &\quad \times \sum_{t=1}^{A/3} \mathbb{P}_*(\tau_n \geq 2A_n/3 \mid Y_t^{t+n-1} \in \mathcal{T}_W(m)) \\ &\geq e^{-n(D_1-D_2)} \text{poly}(n) \\ &\quad \times \sum_{t=1}^{A/3} \mathbb{P}_*(\tau_n \geq 2A_n/3, Y_t^{t+n-1} \in \mathcal{T}_W(m)), \end{aligned}$$

²¹Note that the right-hand side of the first inequality in (36) is well-defined because of (33).

where $D_2 \triangleq D(W\|Q_\star|P_n^2)$, and where the last inequality follows from Fact 3. By summing only over the indices that are multiples of n , we obtain the weaker inequality

$$\Delta_n \geq e^{-n(D_1-D_2)} \text{poly}(n) \times \sum_{j=1}^{A/3n} \mathbb{P}_\star(\tau_n \geq 2A_n/3, Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)). \quad (38)$$

Using (38), we show that $\mathbb{E}(\tau_n - \nu)^+$ grows exponentially with n whenever D_1 and D_2 are both upper bounded by $\alpha - \epsilon$. This, as we saw above, implies that the rate is asymptotically equal to zero, yielding Claim 3.

Let $A = e^{\alpha n}$, and let $\mu \triangleq \epsilon/2$. We rewrite the above summation over $A/3n$ indices as a sum of $A_1 = e^{n(\alpha-D_2-\mu)}/3n$ superblocks of $A_2 = e^{n(D_2+\mu)}$ indices. We have

$$\begin{aligned} \sum_{j=1}^{A/3n} \mathbb{P}_\star(\tau_n \geq 2A_n/3, Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)) \\ = \sum_{s=1}^{A_1} \sum_{j \in I_s} \mathbb{P}_\star(\tau_n \geq 2A_n/3, Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)), \end{aligned}$$

where I_s denotes the s th superblock of A_2 indices. Applying the union bound (in reverse), we see that

$$\begin{aligned} \sum_{s=1}^{A_1} \sum_{j \in I_s} \mathbb{P}_\star(\tau_n \geq 2A_n/3, Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)) \\ \geq \sum_{s=1}^{A_1} \mathbb{P}_\star\left(\tau_n \geq 2A_n/3, \bigcup_{j \in I_s} \{Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)\}\right). \end{aligned}$$

We now show that each term

$$\mathbb{P}_\star(\tau_n \geq 2A_n/3, \bigcup_{j \in I_s} \{Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)\}) \quad (39)$$

in the above summation is large, say greater than $1/2$, by showing that each of them involves the intersection of two large probability events. This, together with (38), implies that

$$\begin{aligned} \Delta_n &= \text{poly}(n) \Omega(e^{n(\alpha-D_1-\mu)}) \\ &\geq \Omega(\exp(n\epsilon/2)) \end{aligned} \quad (40)$$

since $D_1 \leq \alpha - \epsilon$, yielding the desired result.²²

Letting \mathcal{E} denote the decoding error event, we have for all n large enough

$$\begin{aligned} \epsilon &\geq \mathbb{P}_m(\mathcal{E}) \\ &\geq \mathbb{P}_m(\mathcal{E} | \nu > 2A_n/3, \tau_n \leq 2A_n/3) \\ &\quad \times \mathbb{P}_m(\nu > 2A_n/3, \tau_n \leq 2A_n/3) \\ &\geq \frac{1}{2} \mathbb{P}_m(\nu > 2A_n/3) \mathbb{P}_m(\tau_n \leq 2A_n/3 | \nu > 2A_n/3) \\ &\geq \frac{1}{6} \mathbb{P}_m(\tau_n \leq 2A_n/3 | \nu > 2A_n/3). \end{aligned} \quad (41)$$

²²Our proof shows that for all indices n for which $D_1 \leq \alpha - \epsilon$ and $D_2 \leq \alpha - \epsilon$, (40) holds. Therefore, if $D_1 \leq \alpha - \epsilon$ and $D_2 \leq \alpha - \epsilon$ for every n large enough, the reaction delay grows exponentially with n , and thus the rate vanishes. In the case where $D_1 \leq \alpha - \epsilon$ and $D_2 \leq \alpha - \epsilon$ does not hold for all n large enough, but still holds for infinitely many values of n , the corresponding asymptotic rate is still zero by Definition 1.

The third inequality follows by noting that the event $\{\nu > 2A_n/3, \tau_n \leq 2A_n/3\}$ corresponds to the situation where the decoder stops after observing only pure noise. Since a codebook consists of at least two codewords,²³ such an event causes an error with probability at least $1/2$ for at least one message m . Thus, inequality (41) holds under the assumption that m corresponds to such a message.²⁴

Since the event $\{\tau_n \leq 2A_n/3\}$ depends on the channel outputs only up to time $2A_n/3$, we have

$$\mathbb{P}_m(\tau_n \leq 2A_n/3 | \nu > 2A_n/3) = \mathbb{P}_\star(\tau_n \leq 2A_n/3). \quad (42)$$

Combining (42) with (41) we get

$$\mathbb{P}_\star(\tau_n > 2A_n/3) \geq 1 - 6\epsilon. \quad (43)$$

Now, because the Y_{jn}^{jn+n-1} , $j \in I_s$, are i.i.d. under \mathbb{P}_\star ,

$$\begin{aligned} \mathbb{P}_\star\left(\bigcup_{j \in I_s} \{Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)\}\right) \\ = 1 - (1 - \mathbb{P}_\star(Y^n \in \mathcal{T}_W(m)))^{|I_s|}. \end{aligned}$$

From Fact 3 it follows that

$$\mathbb{P}_\star(Y^n \in \mathcal{T}_W(m)) \geq \text{poly}(n) \exp(-nD_2),$$

and by definition $|I_s| = e^{n(D_2+\mu)}$, so

$$\mathbb{P}_\star\left(\bigcup_{j \in I_s} \{Y_{jn}^{jn+n-1} \in \mathcal{T}_W(m)\}\right) = 1 - o(1) \quad (n \rightarrow \infty). \quad (44)$$

Combining (43) and (44), we see that each term (39) involves the intersection of large probability events for at least one message index m . For such a message index, by choosing ϵ sufficiently small, we see that for all sufficiently large n , every single term (39), $s \in \{1, 2, \dots, A_1\}$ is bigger than $1/2$. ■

Finally, to establish the remaining Claim 4, we make use of Theorem 5, whose proof is provided in Appendix B. This theorem implies that any nontrivial codebook contains a (large) set of codewords whose rate is almost the same as the original codebook and whose error probability decays faster than polynomially, say as $e^{-\sqrt{n}}$, with a suitable decoder. Note that we don't use the full implication of Theorem 5.

Proof of Claim 4: The main idea behind the proof is that if Claim 4 does not hold, the noise is likely to produce an output that is "typical" with a codeword before the message is even sent, which means that any decoder must have large error probability. Although the idea is fairly simple, it turns out that a suitable definition for "typical" set and its related error probability analysis make the proof somewhat lengthy.

Proceeding formally, consider inequality (30). This inequality says that, with nonzero probability, the decoder makes a correct decision and stops soon after the beginning of the information transmission period. This motivates the definition of a new random process, which we call the modified output process. With a slight abuse of notation, in the remainder of the proof we use $Y_1, Y_2, \dots, Y_{A+n-1}$ to denote the modified

²³By assumption, see Section III.

²⁴Regarding the fourth inequality in (41), note that $\mathbb{P}_m(\nu > 2A_n/3)$ should be lower bounded by $1/4$ instead of $1/3$ had we taken into account discrepancies due to rounding of noninteger quantities. As mentioned earlier, we disregard these discrepancies as they play no role asymptotically.

output process. The modified output process is generated as if the sent codeword were truncated at the position $\nu + d(n)$, where $d(n)$ is defined in (28). Hence, this process can be thought of as the random process “viewed” by the sequential decoder.

Specifically, the distribution of the modified output process is as follows. If

$$n \geq \lfloor \Delta_n(1 + \epsilon) \rfloor,$$

then the Y_i 's for

$$i \in \{1, \dots, \nu - 1\} \cup \{\nu + \lfloor \Delta_n(1 + \epsilon) \rfloor, \dots, A_n + n - 1\}$$

are i.i.d. according to Q_* , whereas the block

$$Y_\nu, Y_{\nu+1}, \dots, Y_{\nu + \lfloor \Delta_n(1 + \epsilon) \rfloor - 1}$$

is distributed according to $Q(\cdot | c^{d(n)})$, the output distribution given that a *randomly selected* codeword has been transmitted. Note that, in the conditioning, we use $c^{d(n)}$ instead of $c^{d(n)}(m)$ to emphasize that the output distribution is averaged over all possible messages, i.e., by definition

$$Q(y^{d(n)} | c^{d(n)}) = \frac{1}{M} \sum_{m=1}^M Q(y^{d(n)} | c^{d(n)}(m)).$$

Instead, if

$$n < \lfloor \Delta_n(1 + \epsilon) \rfloor,$$

then the modified output process has the same distribution as the original one, i.e., the Y_i 's for

$$i \in \{1, \dots, \nu - 1\} \cup \{\nu + n, \dots, A_n + n - 1\}$$

are i.i.d. according to Q_* , whereas the block

$$Y_\nu, Y_{\nu+1}, \dots, Y_{\nu+n-1}$$

is distributed according to $Q(\cdot | c^n)$.

Consider the following augmented decoder that, in addition to declaring a message, also outputs the time interval

$$[\tau_n - \lfloor \Delta_n(1 + \epsilon) \rfloor + 1, \tau_n - \lfloor \Delta_n(1 + \epsilon) \rfloor + 2, \dots, \tau_n],$$

of size $\lfloor \Delta_n(1 + \epsilon) \rfloor$. A simple consequence of the right-hand side of (30) being (asymptotically) bounded away from zero is that, for n large enough, if the augmented decoder is given a modified output process instead of the original one, with a strictly positive probability it declares the correct message, and the time interval it outputs contains ν .

Now, suppose the decoder is given the modified output process and that it is revealed that the (possibly truncated) sent codeword was sent in one of the

$$r_n = \left\lfloor \frac{(A_n + n - 1) - (\nu \bmod d(n))}{d(n)} \right\rfloor \quad (45)$$

consecutive blocks of duration $d(n)$, as shown in Fig. 12. Using this additional knowledge, the decoder can now both declare the sent message and output a list of

$$\ell_n = \lceil \lfloor \Delta_n(1 + \epsilon) \rfloor / d(n) \rceil \quad (46)$$

block positions, one of which corresponding to the sent message, with a probability strictly away from zero for all n large enough. To do this the decoder, at time τ_n , declares



Fig. 12. Parsing of the entire received sequence of size $A + n - 1$ into r_n blocks of length $d(n)$, one of which being generated by the sent message, and the others being generated by noise.

the decoded message and declares the ℓ_n blocks that overlap with the time indices in

$$\{\tau_n - \lfloor \Delta_n(1 + \epsilon) \rfloor + 1, \tau_n - \lfloor \Delta_n(1 + \epsilon) \rfloor + 2, \dots, \tau_n\}.$$

We now show that the above task that consists of declaring the sent message and producing a list of ℓ_n blocks of size $d(n)$, one of which being the output of the transmitted message, can be performed only if α satisfies Claim 4. To that aim we consider the performance of the (optimal) maximum likelihood decoder that observes output sequences of maximal length

$$d(n) \cdot r_n.$$

Given a sample $y_1, y_2, \dots, y_{A+n-1}$ of the modified output process, and its parsing into consecutive blocks of duration $d(n)$, the optimal decoder outputs a list of ℓ_n blocks that are most likely to occur. More precisely, the maximum likelihood ℓ_n -list decoder operates as follows. For each message m , it finds a list of ℓ_n blocks $y^{d(n)}$ (among all r_n blocks) that maximize the ratio

$$\frac{Q(y^{d(n)} | c^{d(n)}(m))}{Q(y^{d(n)} | \star)}, \quad (47)$$

and computes the sum of these ratios. The maximum likelihood ℓ_n -list decoder then outputs the list whose sum is maximal, and declares the corresponding message.²⁵

The rest of the proof consists in deriving an upper bound on the probability of correct maximum likelihood ℓ_n -list decoding, and show that this bound tends to zero if Claim 4 is not satisfied. To that aim, we first quantify the probability that the noise distribution Q_* outputs a sequence that is typical with a codeword, since the performance of the maximum likelihood ℓ_n -list decoder depends on this probability, as we show below.

By assumption, $(\mathcal{C}'_n, (\tau_n, \phi_n))$ achieves a probability of error $\epsilon'_n \rightarrow 0$ as $n \rightarrow \infty$ at the asynchronism exponent α . This implies that \mathcal{C}'_n can also achieve a nontrivial error probability on the synchronous channel (i.e., with $A = 1$). Specifically, by using the same argument as for (30), we deduce that we can use \mathcal{C}'_n on the synchronous channel, force decoding to happen at the fixed time

$$d(n) = \min\{n, \lfloor (1 + \epsilon)\Delta_n \rfloor\},$$

²⁵To see this, consider a channel output $y^{d(n) \cdot r_n}$ that is composed of r_n consecutive blocks of size $d(n)$, where the j th block is generated by codeword $c^{d(n)}$ and where all the other blocks are generated by noise. The probability of this channel output is

$$\mathbb{P}(y^{d(n) \cdot r_n} | m, j) = Q(y^{d(n)}(j) | c^{d(n)}) \prod_{i \neq j} Q_*(y^{d(n)}(i))$$

where $y^{d(n)}(j)$, $j \in \{1, 2, \dots, r_n\}$, denotes the j th bloc of $y^{d(n) \cdot r_n}$

where Δ_n corresponds to the reaction delay obtained by $(\mathcal{C}'_n, (\tau_n, \phi_n))$ in the asynchronous setting, and guarantee a (maximum) probability of error ϵ''_n such that

$$\epsilon''_n \leq \frac{1}{1+\epsilon} + \epsilon'_n$$

with a suitable decoder. Since the right-hand side of the above inequality is strictly below one for n large enough, Theorem 5 with $q = 1/4$ implies that the code \mathcal{C}'_n has a large subcode $\tilde{\mathcal{C}}_n$, i.e., of almost the same rate with respect to $d(n)$, that, together with an appropriate decoding function $\tilde{\phi}_n$, achieves a maximum error probability at most equal to

$$\epsilon_n = 2(n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|} \exp(-\sqrt{n}/(2 \ln 2)) \quad (48)$$

for all n large enough.

We now start a digression on the code $(\tilde{\mathcal{C}}_n, \tilde{\phi}_n)$ when used on channel Q synchronously. The point is to exhibit a set of “typical output sequences” that cause the decoder $\tilde{\phi}_n$ to make an error with “large probability.” We then move back to the asynchronous channel Q and show that when Claim 4 does not hold, the noise distribution Q_* is likely to produce typical output sequences, thereby inducing the maximum likelihood ℓ_n -list decoder into error.

Unless stated otherwise, we now consider $(\tilde{\mathcal{C}}_n, \tilde{\phi}_n)$ when used on the synchronous channel. In particular error events are defined with respect to this setting.

The set of typical output sequences is obtained through a few steps. We first define the set \mathcal{A}_m with respect to codeword $c^{d(n)}(m) \in \tilde{\mathcal{C}}_n$ as

$$\mathcal{A}_m = \{y^{d(n)} \in \mathcal{T}_W(c^{d(n)}(m)) \text{ with } W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}} : \mathbb{P}(\mathcal{T}_W(c^{d(n)}(m)) | c^{d(n)}(m)) \geq \sqrt{\epsilon_{d(n)}}\} \quad (49)$$

where ϵ_n is defined in (48).

Note that, by using Fact 3, it can easily be checked that \mathcal{A}_m is nonempty for n large enough (depending on $|\mathcal{X}|$ and $|\mathcal{Y}|$), which we assume throughout the argument. For a fixed m , consider the set of sequences in \mathcal{A}_m that maximize (47). These sequences form a set $\mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))$, for some $\tilde{Q} \in \mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}$. It follows that for every message index m for which $c^{d(n)}(m) \in \tilde{\mathcal{C}}_n$, we have

$$\begin{aligned} \epsilon_{d(n)} &\geq \mathbb{P}_m(\mathcal{E}) \\ &\geq \mathbb{P}_m(\mathcal{E} | \{Y_{\nu}^{\nu+d(n)-1} \in \mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))\}) \\ &\times \mathbb{P}_m(\{Y_{\nu}^{\nu+d(n)-1} \in \mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))\}) \\ &\geq \mathbb{P}_m(\mathcal{E} | \{Y_{\nu}^{\nu+d(n)-1} \in \mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))\}) \sqrt{\epsilon_{d(n)}} \\ &\geq \mathbb{P}_m(\mathcal{E} | \{Y_{\nu}^{\nu+d(n)-1} \in \mathcal{B}_m\}) \times \\ &\mathbb{P}_m(\{Y_{\nu}^{\nu+d(n)-1} \in \mathcal{B}_m\} | \{Y_{\nu}^{\nu+d(n)-1} \in \mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))\}) \\ &\times \sqrt{\epsilon_{d(n)}} \\ &\geq \frac{1}{2} \times \\ &\mathbb{P}_m(\{Y_{\nu}^{\nu+d(n)-1} \in \mathcal{B}_m\} | \{Y_{\nu}^{\nu+d(n)-1} \in \mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))\}) \\ &\times \sqrt{\epsilon_{d(n)}} \end{aligned} \quad (50)$$

where for the third inequality we used the definition of \tilde{Q} ; where on the right-hand side of the fourth inequality we

defined the set

$$\mathcal{B}_m \triangleq \{y^{d(n)} \in \mathcal{T}_{\tilde{Q}}(c^{d(n)}(m)) \cap \left(\bigcup_{m' \neq m} \mathcal{T}_{\tilde{Q}}(c^{d(n)}(m')) \right)\};$$

and where the fifth inequality follows from this definition.²⁶

From (50) we get

$$\mathbb{P}_m(\{Y_{\nu}^{\nu+d(n)-1} \in \mathcal{B}_m\} | \{Y_{\nu}^{\nu+d(n)-1} \in \mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))\}) \leq 2\sqrt{\epsilon_{d(n)}}. \quad (51)$$

Therefore, by defining $\tilde{\mathcal{B}}_m$ as

$$\tilde{\mathcal{B}}_m \triangleq \mathcal{T}_{\tilde{Q}}(c^{d(n)}(m)) \setminus \mathcal{B}_m$$

the complement of \mathcal{B}_m in $\mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))$, it follows from (51) that

$$|\tilde{\mathcal{B}}_m| > (1 - 2\sqrt{\epsilon_{d(n)}}) |\mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))|,$$

since under \mathbb{P}_m all the sequences in $\mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))$ are equiprobable.

The set $\bigcup_{m' \neq m}^M \tilde{\mathcal{B}}_{m'}$ is the sought set of “typical output sequences” that causes the decoder make an error with “high probability” conditioned on the sending of message m and conditioned on the channel outputting a sequence in $\mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))$. This ends our digression on $(\tilde{\mathcal{C}}_n, \tilde{\phi}_n)$.

We now compute a lower bound on the probability under Q_* of producing a sequence in $\bigcup_{m=2}^M \tilde{\mathcal{B}}_m$. Because the sets $\{\mathcal{B}_m\}$ are disjoint, we deduce that

$$\begin{aligned} |\bigcup_{m=2}^M \tilde{\mathcal{B}}_m| &\geq (1 - 2\sqrt{\epsilon_n}) \sum_{m=2}^M |\mathcal{T}_{\tilde{Q}}(c^{d(n)}(m))| \\ &\geq \frac{(1 - 2\sqrt{\epsilon_n})}{(d(n) + 1)^{|\mathcal{X}| \cdot |\mathcal{Y}|}} (M - 1) e^{d(n)H(\tilde{Q}|P_n^1)} \\ &\geq \frac{1}{(4n)^{|\mathcal{X}| \cdot |\mathcal{Y}|}} e^{d(n)(H(\tilde{Q}|P_n^1) + \ln M/d(n))} \end{aligned} \quad (52)$$

for all n large enough. For the second inequality we used [7, Lemma 2.5, p. 31]. For the third inequality we used the fact that $d(n) \leq n$, $M \geq 2$, $(1 - 2\sqrt{\epsilon_{d(n)}}) \geq 1/2$ for n large enough,²⁷ and that, without loss of generality, we may assume that $|\mathcal{X}| \cdot |\mathcal{Y}| \geq 2$ since the synchronous capacity C is non-zero—as we assume throughout the paper. Hence we get

$$\begin{aligned} Q_*(\bigcup_{m=2}^M \tilde{\mathcal{B}}_m) &= \sum_{y^{d(n)} \in \bigcup_{m=2}^M \tilde{\mathcal{B}}_m} Q_*(y^{d(n)}) \\ &\geq |\bigcup_{m=2}^M \tilde{\mathcal{B}}_m| \min_{y^{d(n)} \in \bigcup_{m=2}^M \tilde{\mathcal{B}}_m} Q_*(y^{d(n)}) \\ &\geq \frac{1}{(4n)^{|\mathcal{X}| \cdot |\mathcal{Y}|}} e^{d(n)(H(\tilde{Q}|P_n^1) + (\ln M)/d(n))} \\ &\quad \times e^{-d(n)(D((P_n^1 \tilde{Q})_{\mathcal{Y}} \| Q_*) + H((P_n^1 \tilde{Q})_{\mathcal{Y}}))} \end{aligned}$$

²⁶Note that, given that message m is sent, if the channel produces a sequence in \mathcal{B}_m at its output, the (standard) optimal maximum likelihood decoder makes an error with probability at least half. Hence the decoding rule $\tilde{\phi}_n$ also makes an error with probability at least half.

²⁷Note that $d(n) \xrightarrow{n \rightarrow \infty} \infty$ since the coding scheme under consideration achieves a strictly positive rate.

for all n large enough, where for the second inequality we used (52) and [11, Theorem 11.1.2, p. 349]. Letting

$$e_n \triangleq \ln I(P_n^1 \bar{Q}) - (\ln M)/d(n) + D((P_n^1 \bar{Q})_{\mathcal{Y}} \| Q_{\star}),$$

we thus have

$$Q_{\star}(\cup_{m=2}^M \tilde{\mathcal{B}}_m) \geq \frac{1}{(4n)^{2|\mathcal{X}||\mathcal{Y}|}} e^{-e_n \cdot d(n)} \quad (53)$$

for n large enough.

Using (53), we now prove Claim 4 by contradiction. Specifically, assuming that

$$\alpha > \frac{d(n)}{n} e_n + \epsilon/2 \quad \text{for infinitely many indices } n, \quad (54)$$

we prove that, given message $m = 1$ is sent, the probability of error of the maximum likelihood ℓ_n -list decoder does not converge to zero. As final step, we prove that the opposite of (54) implies Claim 4.

Define the events

$$\begin{aligned} \mathcal{E}_1 &= \{Y_{\nu}^{\nu+n-1} \notin \mathcal{A}_1\}, \\ \mathcal{E}_2 &= \{Z \leq \frac{1}{2} \frac{1}{(4n)^{2|\mathcal{X}||\mathcal{Y}|}} e^{\alpha n - e_n \cdot d(n)}\}, \end{aligned}$$

where \mathcal{A}_1 is defined in (49), and where Z denotes the random variable that counts the number of blocks generated by Q_{\star} that are in $\cup_{m=2}^M \tilde{\mathcal{B}}_m$. Define also the complement set

$$\mathcal{E}_3 \triangleq (\mathcal{E}_1 \cup \mathcal{E}_2)^c.$$

The probability that the maximum likelihood ℓ_n -list decoder makes a *correct* decision given that message $m = 1$ is sent is upper bounded as

$$\begin{aligned} \mathbb{P}_1(\mathcal{E}^c) &= \sum_{i=1}^3 \mathbb{P}_1(\mathcal{E}^c | \mathcal{E}_i) \mathbb{P}_1(\mathcal{E}_i) \\ &\leq \mathbb{P}_1(\mathcal{E}_1) + \mathbb{P}_1(\mathcal{E}_2) + \mathbb{P}_1(\mathcal{E}^c | \mathcal{E}_3). \end{aligned} \quad (55)$$

From the definition of \mathcal{A}_1 , we have

$$\mathbb{P}_1(\mathcal{E}_1) = o(1) \quad (n \rightarrow \infty). \quad (56)$$

Now for $\mathbb{P}_1(\mathcal{E}_2)$. There are $r_n - 1$ blocks independently generated by Q_{\star} (r_n is defined in (45)). Each of these blocks has a probability at least equal to the right-hand side of (53) to fall within $\cup_{m=2}^M \tilde{\mathcal{B}}_m$. Hence, using (53) we get

$$\begin{aligned} \mathbb{E}_1 Z &\geq (r_n - 1) \frac{1}{(4n)^{2|\mathcal{X}||\mathcal{Y}|}} e^{-e_n d(n)} \\ &\geq \frac{1}{(4n)^{2|\mathcal{X}||\mathcal{Y}|}} e^{\alpha n - e_n d(n)} \end{aligned} \quad (57)$$

since $r_n \geq e^{\alpha n}/n$. Therefore,

$$\begin{aligned} \mathbb{P}_1(\mathcal{E}_2) &\leq \mathbb{P}_1(Z \leq (\mathbb{E}_1 Z)/2) \\ &\leq \frac{4}{\mathbb{E}_1 Z} \\ &\leq \text{poly}(n) e^{-\alpha n + e_n d(n)} \end{aligned} \quad (58)$$

where the first inequality follows from (57) and the definition of \mathcal{E}_2 ; where for the second inequality we used Chebyshev's inequality and the fact that the variance of a binomial is upper

bounded by its mean; and where for the third inequality we used (57).

Finally for $\mathbb{P}_1(\mathcal{E}^c | \mathcal{E}_3)$. Given \mathcal{E}_3 , the decoder sees at least

$$\frac{1}{2} \frac{1}{(4n)^{2|\mathcal{X}||\mathcal{Y}|}} e^{\alpha n - e_n \cdot d(n)}$$

time slots whose corresponding ratios (47) are at least as large as the one induced by the correct block $Y_{\nu}^{\nu+d(n)-1}$. Hence, given \mathcal{E}_3 , the decoder produces a list of ℓ_n block positions, one of which corresponds to the sent message, with probability at most

$$\begin{aligned} \mathbb{P}_1(\mathcal{E}^c | \mathcal{E}_3) &\leq \ell_n \left(\frac{1}{2} \frac{1}{(4n)^{2|\mathcal{X}||\mathcal{Y}|}} e^{\alpha n - e_n \cdot d(n)} \right)^{-1} \\ &= \text{poly}(n) e^{-\alpha n + e_n \cdot d(n)}, \end{aligned} \quad (59)$$

where the first inequality follows from union bound, and where for the equality we used the fact that finite rate implies $\ell_n = \text{poly}(n)$.²⁸

From (55), (56), (58), and (59), the probability that the maximum likelihood ℓ_n -list decoder makes a correct decision, $\mathbb{P}_1(\mathcal{E}^c)$, is arbitrarily small for infinitely many indices n whenever (54) holds. Therefore to achieve vanishing error probability we must have, for all n large enough,

$$\begin{aligned} \alpha &\leq \frac{d(n)}{n} (I(P_n^1 \bar{Q}) - (\ln M)/d(n) + D((P_n^1 \bar{Q})_{\mathcal{Y}} \| Q_{\star})) \\ &\quad + \epsilon/2. \end{aligned} \quad (60)$$

We now show, via a continuity argument, that the above condition implies Claim 4. Recall that $\bar{Q} \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, defined just after (49), depends on n and has the property

$$\mathbb{P}(\mathcal{T}_{\bar{Q}}(c^{d(n)}(m) | c^{d(n)}(m))) \geq \sqrt{\epsilon_{d(n)}}. \quad (61)$$

Now, from Fact 3 we also have the upper bound

$$\mathbb{P}(\mathcal{T}_{\bar{Q}}(c^{d(n)}(m) | c^{d(n)}(m))) \leq e^{-d(n)D(\bar{Q} \| Q | P_n^1)}. \quad (62)$$

Since $\sqrt{\epsilon_{d(n)}} = \Omega(e^{-\sqrt{d(n)}})$, from (61) and (62) we get

$$D(\bar{Q} \| Q | P_n^1) \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and therefore

$$\|P_n^1 \bar{Q} - P_n^1 Q\| \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

where $\|\cdot\|$ denotes the L_1 norm. Hence, by continuity of the divergence, condition (60) gives, for all n large enough,

$$\begin{aligned} \alpha &\leq \frac{d(n)}{n} (I(P_n^1 Q) - (\ln M)/d(n) + D((P_n^1 Q)_{\mathcal{Y}} \| Q_{\star})) \\ &\quad + \epsilon \end{aligned} \quad (63)$$

which yields Claim 4. \blacksquare

²⁸This follows from the definition of rate $R = \ln M/\mathbb{E}(\tau - \nu)^+$, the fact that $\ln M/n \leq C$ for reliable communication, and the definition of ℓ_n (46).

C. Proof of Corollary 3

By assumption α_o is nonzero since divergence is always non-negative. This implies that the synchronous capacity is nonzero by the last claim of Theorem 1. This, in turn, implies that (R, α) is achievable for some sufficiently small $R > 0$ and $\alpha > 0$ by [3, Corollary 1].

Using Theorem 3,

$$\alpha \leq \alpha(R) \leq \max_s \alpha_2 \quad (64)$$

where α_2 is given by expression (10). In this expression, by letting $W = Q_*$ in the minimization, we deduce that $\alpha_2 \leq D(Q_* \| Q | P_2)$, and therefore

$$\begin{aligned} \max_s \alpha_2 &\leq \max_s D(Q_* \| Q | P_2) \\ &= \max_{P_2} D(Q_* \| Q | P_2) \\ &= \max_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} Q_*(y) \ln \frac{Q_*(y)}{Q(y|x)} \\ &= \max_{x \in \mathcal{X}} D(Q_* \| Q(\cdot|x)), \end{aligned}$$

and from (64) we get

$$\alpha \leq \max_{x \in \mathcal{X}} D(Q_* \| Q(\cdot|x)).$$

Since, by assumption,

$$\alpha_o > \max_{x \in \mathcal{X}} D(Q_* \| Q(\cdot|x)),$$

and since $\alpha_o = \alpha(R=0)$ by Theorem 1, it follows that $\alpha(R)$ is discontinuous at $R=0$. ■

D. Proof of Theorem 4

We first exhibit a coding scheme that achieves any (R, α) with $R \leq C$ and

$$\alpha \leq \max_{P \in \mathcal{P}^{\mathcal{X}}} \min_{W \in \mathcal{P}^{\mathcal{Y}|X}} \max\{D(W \| Q | P), D(W \| Q_* | P)\}.$$

All codewords start with a common preamble that is composed of $(\ln(n))^2$ repetitions of a symbol x such that $D(Q(\cdot|x) \| Q_*) = \infty$ (such a symbol exists since $\alpha_o = \infty$). The next $(\ln(n))^3$ symbols of each codeword are drawn from a code that achieves a rate equal to $R - \epsilon$ on the synchronous channel. Finally, all the codewords end with a common large suffix s^l of size $l = n - (\ln(n))^2 - (\ln(n))^3$ that has an empirical type P such that, for all $W \in \mathcal{P}^{\mathcal{Y}|X}$, at least one of the following two inequalities holds:

$$\begin{aligned} D(W \| Q | P) &\geq \alpha \\ D(W \| Q_* | P) &\geq \alpha. \end{aligned}$$

The receiver runs two sequential decoders in parallel, and makes a decision whenever one of the two decoder declares a message. If the two decoders declare different messages at the same time, the receiver declares one of the messages at random.

The first decoder tries to identify the sent message by first locating the preamble. At time t it checks if the channel output y_t can be generated by x but cannot be generated by noise, i.e., if

$$Q(y_t|x) > 0 \quad \text{and} \quad Q(y_t|\star) = 0. \quad (65)$$

If condition (65) does not hold, the decoder moves one-step ahead and checks condition (65) at time $t+1$. If condition (65) does hold, the decoder marks the current time as the beginning of the “decoding window” and proceeds to the second step. The second step consists in exactly locating and identifying the sent codeword. Once the beginning of the decoding window has been marked, the decoder makes a decision the first time it observes $(\ln n)^3$ symbols that are typical with one of the codewords. If no such time is found within $(\ln(n))^2 + (\ln(n))^3$ time steps from the time the decoding window has been marked, the decoder declares a random message.

The purpose of the second decoder is to control the average reaction delay by stopping the decoding process in the rare event when the first decoder misses the codeword. Specifically, the second “decoder” is only a stopping rule based on the suffix s^l . At each time t the second decoder checks whether $D(\hat{P}_{Y_{t-l+1}^t} \| Q | P) < \alpha$. If so, the decoder stops and declares a random message. If not, the decoder moves one step ahead.

The arguments for proving that the coding scheme described above achieves (R, α) provided

$$\alpha \leq \max_P \min_W \max\{D(W \| Q | P), D(W \| Q_* | P)\} \quad (66)$$

closely parallel those used to prove Theorem 2, and are therefore omitted.²⁹

The converse is the second part of Theorem 3. ■

E. Proof of Theorem 6

1) *Lower bound:* To establish the lower bound in Theorem 6, we exhibit a training based scheme with preamble size ηn with

$$\eta = (1 - R/C), \quad (67)$$

and that achieves any rate asynchronism pair (R, α) such that

$$\alpha \leq m_1 \left(1 - \frac{R}{C}\right) \quad R \in (0, C] \quad (68)$$

where

$$m_1 \triangleq \max_{P \in \mathcal{P}^{\mathcal{X}}} \min_{W \in \mathcal{P}^{\mathcal{Y}|X}} \max\{D(W \| Q | P), D(W \| Q_* | P)\}.$$

Fix $R \in (0, C]$ and let α satisfy (68). Each codeword starts with a common preamble of size ηn where η is given by (67) and whose empirical distribution is equal to³⁰

$$P_p \triangleq \arg \max_{P \in \mathcal{P}^{\mathcal{X}}} \left(\min_{W \in \mathcal{P}^{\mathcal{Y}|X}} \max\{D(W \| Q | P), D(W \| Q_* | P)\} \right).$$

The remaining $(1 - \eta)n$ symbols of each codeword are i.i.d. generated according to a distribution P that almost achieves capacity of the synchronous channel, i.e., such that $I(PQ) = C - \epsilon$ for some small $\epsilon > 0$.

Note that by (68) and (67), α is such that for any $W \in \mathcal{P}^{\mathcal{Y}|X}$ at least one of the following two inequalities holds:

$$\begin{aligned} D(W \| Q | P_p) &\geq \alpha/\eta \\ D(W \| Q_* | P_p) &\geq \alpha/\eta. \end{aligned} \quad (69)$$

²⁹In particular, note that the first decoder never stops before time ν .

³⁰ P_p need not be a valid type for finite values of n , but this small discrepancy plays no role asymptotically since P_p can be approximated arbitrarily well with types of order sufficiently large.

The preamble detection rule is to stop the first time when last ηn output symbols $Y_{t-\eta n+1}^t$ induce an empirical conditional probability $\hat{P}_{Y_{t-\eta n+1}^t|x^{\eta n}}$ such that

$$D(\hat{P}_{Y_{t-\eta n+1}^t|x^{\eta n}}||Q|P_p) \leq D(\hat{P}_{Y_{t-\eta n+1}^t|x^{\eta n}}||Q_*|P_p) \quad (70)$$

where $x^{\eta n}$ is the preamble.

When the preamble is located, the decoder makes a decision on the basis of the upcoming $(1-\eta)n$ output symbols using maximum likelihood decoding. If no preamble has been located by time $A_n + n - 1$, the decoder declares a message at random.

We compute the reaction delay and the error probability. For notational convenience, instead of the decoding time, we consider the time τ_n that the decoder detects the preamble, i.e., the first time t such that (70) holds. The actual decoding time occurs $(1-\eta)n$ time instants after the preamble has been detected, i.e., at time $\tau_n + (1-\eta)n$.

For the reaction delay we have

$$\begin{aligned} \mathbb{E}(\tau_n - \nu)^+ &= \mathbb{E}_1(\tau_n - \nu)^+ \\ &= \mathbb{E}_1[(\tau_n - \nu)^+ \mathbb{1}(\tau_n \geq \nu + \eta n)] \\ &\quad + \mathbb{E}_1[(\tau_n - \nu)^+ \mathbb{1}(\tau_n \leq \nu + \eta n - 1)] \\ &\leq (A_n + n - 1) \mathbb{P}_1(\tau_n \geq \nu + \eta n) + \eta n \end{aligned} \quad (71)$$

where, as usual, the subscript 1 in \mathbb{E}_1 and \mathbb{P}_1 indicates conditioning on the event that message $m = 1$ is sent. A similar computation as in (20) yields

$$\begin{aligned} \mathbb{P}_1(\tau_n \geq \nu + \eta n) &\leq \mathbb{P}_1(D(\hat{P}_{Y_{\nu+\eta n-1}^{\nu+\eta n-1}|x^{\eta n}}||Q|P_p) \geq \alpha/\eta) \\ &\leq \sum_{W \in \mathcal{P}_n^{\mathcal{Y}|X}: D(W||Q|P_p) \geq \alpha/\eta} e^{-\eta n D(W||Q|P_p)} \\ &\leq \text{poly}(n) e^{-n\alpha}. \end{aligned} \quad (72)$$

The first inequality follows from the fact that event $\{\tau_n \geq \nu + n\}$ is included into event

$$\{D(\hat{P}_{Y_{\nu+\eta n-1}^{\nu+\eta n-1}|x^{\eta n}}||Q|P_p) > D(\hat{P}_{Y_{\nu+\eta n-1}^{\nu+\eta n-1}|x^{\eta n}}||Q_*|P_p)\}$$

which, in turn, is included into event

$$\{D(\hat{P}_{Y_{\nu+\eta n-1}^{\nu+\eta n-1}|x^{\eta n}}||Q|P_p) \geq \alpha/\eta\}$$

because of (69). The second inequality follows from Fact 2. Hence, from (71) and (72)

$$\mathbb{E}(\tau_n - \nu)^+ \leq \eta n + o(1) \quad (73)$$

whenever $A_n = e^{n(\alpha-\epsilon)}$, $\epsilon > 0$. Since the actual decoding time occurs $(1-\eta)n$ time instants after τ_n , where $\eta = (1-R/C)$, and that the code used to transmit information achieves the capacity of the synchronous channel, the above strategy operates at rate R .

To show that the above strategy achieves vanishing error probability, one uses arguments similar to those used to prove Theorem 2 (see from paragraph after (21) onwards), so the proof is omitted. There is one little caveat in the analysis that concerns the event when the preamble is located somewhat earlier than its actual timing, i.e., when the decoder locates the preamble over a time period $[t - \eta n + 1, \dots, t]$ with $\nu \leq t \leq$

$\nu + \eta n - 2$. One way to make the probability of this event vanish as $n \rightarrow \infty$, is to have the preamble have a ‘‘sufficiently large’’ Hamming distance with any of its shifts. To guarantee this, one just needs to modify the original preamble in a few (say, $\log n$) positions. This modifies the preamble type negligibly. For a detailed discussion on how to make this modification, we refer the reader to [9], where the problem is discussed in the context of sequential frame synchronization.

Each instance of the above random coding strategy satisfies the conditions of Definition 3; there is a common preamble of size ηn and the decoder decides to stop at any particular time t based on $Y_{t-n+1}^{t-n+\eta n}$. We now show that there exists a particular instance yielding the desired rate and error probability.

First note that the above rate analysis only depends on the preamble, and not on the codebook that follows the preamble. Hence, because the error probability, averaged over codebooks and messages, vanishes, we deduce that there exists at least one codebook that achieves rate R and whose average over messages error probability tends to zero.

From this code, we remove codewords with poor error probability, say whose error probabilities are at least twice the average error probability. The resulting expurgated code has a rate that tends to R and a vanishing maximum error probability.

2) *Upper bound:* To establish the upper bound it suffices to show that for training based schemes (R, α) with $R > 0$ must satisfy

$$\alpha \leq m_2 \left(1 - \frac{R}{C}\right). \quad (74)$$

The upper bound in Theorem 6 then follows from (74) and the general upper bound derived in Theorem 3.

The upper bound (74) follows from the following lemma:

Lemma 1: A rate $R > 0$ coding scheme whose decoder operates according to a sliding window stopping rule with window size ηn cannot achieve an asynchronism exponent larger than ηm_2 .

Lemma 1 says that any coding scheme with a limited memory stopping rule capable of processing only ηn symbols at a time achieves an asynchronism exponent at most $O(\eta)$, unless $R = 0$ or if the channel is degenerate, i.e., $\alpha_o = m_2 = \infty$, in which case Lemma 1 is trivial and we have the asynchronous capacity expression given by Theorem 4.

To deduce (74) from Lemma 1, consider a training-based scheme which achieves a delay Δ with a non-trivial error probability (i.e., bounded away from 0). Because the preamble conveys no information, the rate is at most

$$C \frac{\min\{\Delta, n\} - \eta n}{\Delta} \leq C(1 - \eta)$$

by the channel coding theorem for a synchronous channel. Hence, for a rate $R > 0$ training-based scheme the training fraction η is upper bounded as

$$\eta \leq 1 - \frac{R}{C}.$$

This implies (74) by Lemma 1. ■

Proof of Lemma 1: The lemma holds trivially if $m_2 = \infty$. We thus assume that $m_2 < \infty$. Consider a training-based

scheme $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}_{n \geq 1}$ in the sense of Definition 3. For notational convenience, we consider τ_n to be the time when the decoder detects the preamble. The actual decoding time (in the sense of Definition 3 part 2) occurs $(1 - \eta)n$ times instants after the preamble has been detected, i.e., at time $\tau_n + (1 - \eta)n$. This allows us to write τ_n as

$$\tau_n = \inf\{t \geq 1 : S_t = 1\},$$

where

$$S_t = S_t(Y_{t-\eta n+1}^t) \quad 1 \leq t \leq A_n + n - 1,$$

referred to as the “stopping rule at time t ,” is a binary random variable such that $\{S_t = 1\}$ represents the set of output sequences $y_{t-\eta n+1}^t$ which make τ_n stop at time t , assuming that τ_n hasn’t stopped before time t .

Now, every sequence $y^{\eta n} \in \mathcal{Y}^{\eta n}$ satisfies

$$Q_*(y^{\eta n}) \geq e^{-m_2 \eta n}.$$

Therefore, any deterministic stopping rule stops at any particular time either with probability zero or with probability at least $e^{-m_2 \eta n}$, i.e., for all t , either the stopping rule S_t satisfies $\mathbb{P}(S_t = 1) \geq e^{-m_2 \eta n}$ or it is trivial in the sense that $\mathbb{P}(S_t = 1) = 0$. For now, we assume that the stopping rule is deterministic; the randomized case follows easily as we describe at the end of the proof.

Let \mathcal{S} denote the subset of indices $t \in \{1, 2, \dots, A_n/4\}$ such that S_t is non-trivial, and let $\bar{\mathcal{S}}_k$ denote the subset of indices in \mathcal{S} that are congruent to $k \bmod \eta n$, i.e.,

$$\bar{\mathcal{S}}_k = \{t : t \in \mathcal{S}, t = j \cdot \eta n + k, j = 0, 1, \dots\}.$$

Note that for each k , the set of stopping rules S_t , $t \in \bar{\mathcal{S}}_k$ are independent since S_t depends only on $Y_{t-\eta n+1}^t$.

By repeating the same argument as in (41)-(42), for any $\epsilon > 0$, for all n large enough and any message index m the error probability $\mathbb{P}_m(\mathcal{E})$ satisfies

$$\begin{aligned} \epsilon &\geq \mathbb{P}_m(\mathcal{E}) \\ &\geq \frac{1}{4} \mathbb{P}_*(\tau_n \leq A_n/2). \end{aligned} \quad (75)$$

Since $\epsilon > 0$ is arbitrary, we deduce

$$\mathbb{P}_*(\tau_n \geq A_n/2) \geq 1/2 \quad (76)$$

i.e., a coding scheme achieves a vanishing error probability only if the probability of stopping after time $A_n/2$ is at least 0.5 when the channel input is all \star 's. Thus, assuming that our coding scheme achieves vanishing error probability, we have

$$|\mathcal{S}| < \eta n e^{m_2 \eta n}.$$

To see this, note that if $|\mathcal{S}| \geq \eta n e^{m_2 \eta n}$, then there exists a value k^* such that $|\bar{\mathcal{S}}_{k^*}| \geq e^{m_2 \eta n}$, and hence

$$\begin{aligned} \mathbb{P}_*(\tau_n \geq A_n/2) &\leq \mathbb{P}_*(S_t = 0, t \in \mathcal{S}) \\ &\leq \mathbb{P}_*(S_t = 0, t \in \bar{\mathcal{S}}_{k^*}) \\ &= (1 - e^{-m_2 \eta n})^{|\bar{\mathcal{S}}_{k^*}|} \\ &\leq (1 - e^{-m_2 \eta n}) e^{m_2 \eta n}. \end{aligned}$$

Since the above last term tends to $1/e < 1/2$ for n large enough, $\mathbb{P}_*(\tau_n \geq A_n/2) < 1/2$ for n large enough, which is in

conflict with the assumption that the coding scheme achieves vanishing error probability.

The fact that $|\mathcal{S}| < \eta n e^{m_2 \eta n}$ implies, as we shall prove later, that

$$\mathbb{P}(\tau_n \geq A_n/2 | \nu \leq A_n/4) \geq \frac{1}{2} \left(1 - \frac{8\eta^2 n^2 e^{m_2 \eta n}}{A_n}\right). \quad (77)$$

Hence,

$$\begin{aligned} \mathbb{E}(\tau_n - \nu)^+ &\geq \mathbb{E}((\tau_n - \nu)^+ | \tau_n \geq A_n/2, \nu \leq A_n/4) \\ &\quad \times \mathbb{P}(\tau_n \geq A_n/2, \nu \leq A_n/4) \\ &\geq \frac{A_n}{16} \mathbb{P}(\tau_n \geq A_n/2 | \nu \leq A_n/4) \\ &\geq \frac{A_n}{32} \left(1 - \frac{8\eta^2 n^2 e^{m_2 \eta n}}{A_n}\right). \end{aligned} \quad (78)$$

where for the second inequality we used the fact that ν is uniformly distributed, and where the third inequality holds by (77). Letting $A_n = e^{\alpha n}$, from (78) we deduce that if $\alpha > m\eta$, then $\mathbb{E}(\tau_n - \nu)^+$ grows exponentially with n , implying that the rate is asymptotically zero.³¹ Hence a sliding window stopping rule which operates on a window of size ηn cannot accommodate a positive rate while achieving an asynchronism exponent larger than ηm . This establishes the desired result.

We now show (77). Let \mathcal{N} be the subset of indices in $\{1, 2, \dots, A_n/4\}$ with the following property. For any $t \in \mathcal{N}$, the $2n$ indices $\{t, t+1, \dots, t+2n-1\}$ do not belong to \mathcal{S} , i.e., all $2n$ of the associated stopping rules are trivial. Then we have

$$\begin{aligned} \mathbb{P}(\tau_n \geq A_n/2 | \nu \leq A_n/4) &\geq \mathbb{P}(\tau_n \geq A_n/2 | \nu \in \mathcal{N}) \\ &\quad \times \mathbb{P}(\nu \in \mathcal{N} | \nu \leq A_n/4) \\ &= \mathbb{P}(\tau_n \geq A_n/2 | \nu \in \mathcal{N}) \frac{|\mathcal{N}|}{A_n/4} \end{aligned} \quad (79)$$

since ν is uniformly distributed. Using that $|\mathcal{S}| < \eta n e^{m_2 \eta n}$,

$$|\mathcal{N}| \geq (A_n/4 - 2\eta n^2 e^{m_2 \eta n}),$$

hence from (79)

$$\begin{aligned} \mathbb{P}(\tau_n \geq A_n/2 | \nu \leq A_n/4) &\geq \mathbb{P}(\tau_n \geq A_n/2 | \nu \in \mathcal{N}) \left(1 - \frac{8\eta^2 n^2 e^{m_2 \eta n}}{A_n}\right). \end{aligned} \quad (80)$$

Now, when $\nu \in \mathcal{N}$, all stopping times that could potentially depend on the transmitted codeword symbols are actually trivial, so the event $\{\tau_n \geq A_n/2\}$ is independent of the symbols sent at times $\nu, \nu+1, \dots, \nu+N-1$. Therefore,

$$\mathbb{P}(\tau_n \geq A_n/2 | \nu \in \mathcal{N}) = \mathbb{P}_*(\tau_n \geq A_n/2). \quad (81)$$

Combining (81) with (80) gives the desired claim (77).

Finally, to see that randomized stopping rules also cannot achieve asynchronism exponents larger than ηm , note that a randomized stopping rule can be viewed as simply a probability distribution over deterministic stopping rules. The previous analysis shows that for any deterministic stopping

³¹Any coding scheme that achieves vanishing error probability cannot have $\ln M$ grow faster than linearly with n , because of the limitation imposed by the capacity of the synchronous channel. Hence, if $\mathbb{E}(\tau_n - \nu)^+$ grows exponentially with n , the rate goes to zero exponentially with n .

rule, and any asynchronism exponent larger than ηm , either the probability of error is large (e.g., at least $1/8$), or the expected delay is exponential in n . Therefore, the same holds for randomized stopping rules. ■

F. Comments on Error Criteria

We end this section by commenting on maximum versus average rate/error probability criteria. The results in this paper consider the rate defined with respect to maximum (over messages) reaction delay and consider maximum (over messages) error probability. Hence all the achievability results also hold when delay and error probability are averaged over messages.

To see that the converse results in this paper also hold for the average case, we use the following standard expurgation argument. Assume $\{(\mathcal{C}_n, (\tau_n, \phi_n))\}$ is an (R, α) coding scheme where the error probability and the delay of $(\mathcal{C}_n, (\tau_n, \phi_n))$ are defined as

$$\epsilon_n \triangleq \frac{1}{M} \sum_{m=1}^M \mathbb{P}_m(\mathcal{E}),$$

and

$$\bar{\Delta}_n \triangleq \frac{1}{M} \sum_{m=1}^M \mathbb{E}_m(\tau_n - \nu)^+,$$

respectively. By definition of an (R, α) coding scheme, this means that given some arbitrarily small $\epsilon > 0$, and for all n large enough,

$$\epsilon_n \leq \epsilon$$

and

$$\frac{\ln M}{\bar{\Delta}_n} \geq R - \epsilon.$$

Hence, for n large enough and any $\delta > 1$, one can find a (nonzero) constant fraction of codewords $\mathcal{C}_n' \subset \mathcal{C}_n$ (\mathcal{C}_n' is the “expurgated” ensemble) that satisfies the following property: the rate defined with respect to maximum (over \mathcal{C}_n') delay is at least $(R - \epsilon)/\delta$ and the maximum error probability is less than $\eta\epsilon$, where $\eta = \eta(\delta) > 0$. One then applies the converse results to the expurgated ensemble to derive bounds on $(R/\delta, \alpha)$, and thus on (R, α) , since $\delta > 1$ can be chosen arbitrarily.

VI. CONCLUDING REMARKS

We analyzed a model for asynchronous communication which captures the situation when information is emitted infrequently. General upper and lower bounds on capacity were derived, which coincide in certain cases. The forms of these bounds are similar and have two parts: a mutual information part and a divergence part. The mutual information part is reminiscent of synchronous communication: to achieve a certain rate, there must be, on average, enough mutual information between the time information is sent and the time it is decoded. The divergence part is novel, and comes from asynchronism. Asynchronism introduces two additional error events that must be overcome by the decoder. The first event happens when the noise produces a channel output that looks as if it was generated by a codeword. The larger the level of asynchronism, the more likely this event becomes. The second event happens when the channel behaves atypically,

which results in the decoder missing the codeword. When this event happens, the rate penalty is huge, on the order of the asynchronism level. As such, the second event contributes to increased average reaction delay, or equivalently, lowers the rate. The divergence part in our upper and lower bounds on capacity strikes a balance between these two events.

An important conclusion of our analysis is that, in general, training-based schemes are not optimal in the high rate, high asynchronism regime. In this regime, training-based architectures are unreliable, whereas it is still possible to achieve an arbitrarily low probability of error using strategies that combine synchronization with information transmission.

Finally, we note that further analysis is possible when we restrict attention to a simpler slotted communication model in which the possible transmission slots are nonoverlapping and contiguous. In particular, for this more constrained model [13] develops a variety of results, among which is that except in somewhat pathological cases, training-based schemes are strictly suboptimal at all rates below the synchronous capacity. Additionally, the performance gap is quantified for the special cases of the binary symmetric and additive white Gaussian noise channels, where it is seen to be significant in the high rate regime but vanish in the limit of low rates. Whether the characteristics observed for the slotted model are also shared by unslotted models remains to be determined, and is a natural direction for future research.

ACKNOWLEDGMENTS

The authors are grateful to the reviewers for their insightful and detailed comments which very much contributed to improve the paper. The authors would also like to thank the associate editors Suhas Diggavi and Tsachy Weissman and the editor-in-chief Helmut Bölcskei for their care in handling this paper. This paper also benefited from useful discussions with Sae-Young Chung and Da Wang.

APPENDIX A PROOF OF REMARK 2 (P. 14)

To show that the random coding scheme proposed in the proof of Theorem 2 achieves (6) with equality, we show that

$$\alpha \leq \max_{P: I(PQ) \geq R} \min_{V \in \mathcal{P}^{\mathcal{Y}}} \max\{D(V \| (PQ)_{\mathcal{Y}}), D(V \| Q_{\star})\}. \quad (82)$$

Recall that, by symmetry of the encoding and decoding procedures, the average reaction delay is the same for any message. Hence

$$\Delta_n = \mathbb{E}_1(\tau_n - \nu)^+,$$

where \mathbb{E}_1 denotes expectation under the probability measure \mathbb{P}_1 , the channel output distribution when message 1 is sent, averaged over time and codebooks.

Suppose for the moment that

$$\mathbb{E}_1(\tau_n - \nu)^+ \geq n(1 - o(1)) \quad n \rightarrow \infty. \quad (83)$$

It then follows from Fano's inequality that the input distribution P must satisfy $I(PQ) \geq R$. Hence, to establish (82) we will show that at least one of the following inequalities

$$\begin{aligned} D(V\|(PQ)_y) &\geq \alpha \\ D(V\|Q_*) &\geq \alpha \end{aligned} \quad (84)$$

holds for any $V \in \mathcal{P}^y$. The arguments are similar to those used to establish Claim 3 of Theorem 3. Below we provide the key steps.

We proceed by contradiction and show that if both the inequalities in (84) are reversed, then the asymptotic rate is zero. To that aim we provide a lower bound on $\mathbb{E}_1(\tau_n - \nu)^+$.

Let τ'_n denote the time of the beginning of the decoding window, i.e., the first time when the previous n output symbols have empirical distribution \hat{P} such that $D(\hat{P}\|Q_*) \geq \alpha$. By definition, $\tau_n \geq \tau'_n$, so

$$\begin{aligned} \mathbb{E}_1(\tau_n - \nu)^+ &\geq \mathbb{E}_1(\tau'_n - \nu)^+ \\ &\geq \frac{1}{3} \sum_{t=1}^{A/3} \mathbb{P}_{1,t}(\tau'_n \geq 2A_n/3), \end{aligned} \quad (85)$$

where the second inequality follows from Markov's inequality, and where $\mathbb{P}_{1,t}$ denotes the probability measure at the output of the channel conditioned on the event that message 1 starts being sent at time t , and averaged over codebooks. Note that, because τ'_n is not a function of the codebook, there is no averaging on the stopping times.³²

Fix $V \in \mathcal{P}_y$. We lower bound each term $\mathbb{P}_{1,t}(\tau'_n \geq 2A_n/3)$ in the above sum as

$$\begin{aligned} \mathbb{P}_{1,t}(\tau'_n \geq 2A_n/3) &\geq \mathbb{P}_{1,t}(\tau'_n \geq 2A_n/3 | Y_t^{t+n-1} \in \mathcal{T}_V) \mathbb{P}_{1,t}(Y_t^{t+n-1} \in \mathcal{T}_V) \\ &\geq \mathbb{P}_{1,t}(\tau_n \geq 2A_n/3 | Y_t^{t+n-1} \in \mathcal{T}_V) e^{-nD_1} \text{poly}(n), \end{aligned} \quad (86)$$

where $D_1 \triangleq D(V\|(PQ)_y)$, and where the second inequality follows from Fact 2.

The key change of measure step (37) results now in the equality

$$\begin{aligned} \mathbb{P}_{1,t}(\tau'_n \geq 2A_n/3 | Y_t^{t+n-1} \in \mathcal{T}_V) &= \mathbb{P}_*(\tau'_n \geq 2A_n/3 | Y_t^{t+n-1} \in \mathcal{T}_V), \end{aligned} \quad (87)$$

which can easily be checked by noticing that the probability of any sequence y_t^{t+n-1} in \mathcal{T}_V is the same under $\mathbb{P}_{1,t}$. Substituting (87) into the right-hand side of (86), and using (85) and Fact 2, we get

$$\begin{aligned} \mathbb{E}_1(\tau_n - \nu)^+ &\geq e^{-n(D_1 - D_2)} \text{poly}(n) \\ &\quad \times \sum_{t=1}^{A/3} \mathbb{P}_*(\tau_n \geq 2A_n/3, Y_t^{t+n-1} \in \mathcal{T}_V), \end{aligned} \quad (88)$$

where $D_2 \triangleq D(V\|Q_*)$. The rest of the proof consists in showing that if the two inequalities in (84) are reversed, then the right-hand side of the above inequality grows exponentially with n , which results in an asymptotic rate equal to zero.

³²For different codebook realizations, stopping rule τ'_n is the same, by contrast with τ_n which depends on the codebook via the joint typicality criterion of the second phase.

The arguments closely parallel the ones that prove Claim 3 of Theorem 3 (see from (38) onwards), and hence are omitted.

To conclude the proof we show (83). Using the alternate form of expectation for non-negative random variables $\mathbb{E}X = \sum_{k \geq 0} \mathbb{P}(X \geq k)$, we have

$$\begin{aligned} \mathbb{E}_1(\tau_n - \nu)^+ &\geq \sum_{i=1}^{g(n)} \mathbb{P}_1(\tau_n \geq \nu + i) \\ &\geq \sum_{i=1}^{g(n)} (1 - \mathbb{P}_1(\tau_n < \nu + i)) \\ &\geq g(n)(1 - \mathbb{P}_1(\tau_n \leq \nu + g(n))), \end{aligned}$$

where we defined

$$g(n) \triangleq n - \lceil n^{3/4} \rceil,$$

and where the last inequality follows from the fact that $\mathbb{P}_1(\tau_n < \nu + i)$ is a non-decreasing function of i . Since $g(n) = n(1 - o(1))$, to establish (83) it suffices to show that

$$\mathbb{P}_1(\tau_n \leq \nu + g(n)) = o(1) \quad (n \rightarrow \infty). \quad (89)$$

Since

$$\mathbb{P}_1(\tau_n < \nu) = o(1) \quad (n \rightarrow \infty),$$

as follows from computation steps in (22) and (23), to establish (89) it suffices to show that

$$\mathbb{P}_1(\nu \leq \tau_n \leq \nu + g(n)) = o(1) \quad (n \rightarrow \infty). \quad (90)$$

For $i \in \{0, 1, \dots, g(n)\}$ we have

$$\begin{aligned} \mathbb{P}_1(\tau_n = \nu + i) &\leq \mathbb{P}_1\left(\|\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} PQ\| \leq \mu \cdot |\mathcal{X}| \cdot |\mathcal{Y}|\right) \\ &= \sum_J \mathbb{P}_1\left(\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} = J\right) \end{aligned} \quad (91)$$

where the above summation is over all typical joint types, i.e., all $J \in \mathcal{P}_{\mathcal{X}, \mathcal{Y}}^n$ such that

$$|\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}}(a, b) - J(a, b)| \leq \mu \quad (92)$$

for all $(a, b) \in \mathcal{X} \times \mathcal{Y}$.

We upper bound each term in this summation. First observe that event

$$\{\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} = J\},$$

for $i \in \{0, 1, \dots, g(n)\}$, involves random vector $Y_{\nu+i-n+1}^{\nu+i}$ which is partly generated by noise and partly generated by the transmitted codeword corresponding to message 1. In the following computation k refers to first symbols of $Y_{\nu+i-n+1}^{\nu+i}$ which are generated by noise, i.e., by definition $k = n - (i+1)$. Note that since $0 \leq i \leq g(n)$, we have

$$\lceil n^{3/4} \rceil - 1 \leq k \leq n - 1.$$

We have

$$\begin{aligned} \mathbb{P}_1(\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} = J) \\ = \sum_{\substack{J_1 \in \mathcal{P}_k \\ J_2 \in \mathcal{P}_{n-k} \\ kJ_1 + (n-k)J_2 = nJ}} \left(\sum_{(x^k, y^k): \hat{P}_{x^k, y^k} = J_1} P(x^k) Q_*(y^k) \right) \\ \times \left(\sum_{(x^{n-k}, y^{n-k}): \hat{P}_{x^{n-k}, y^{n-k}} = J_2} \mathbb{P}(x^{n-k}, y^{n-k}) \right), \quad (93) \end{aligned}$$

where we used the following shorthand notations for probabilities

$$\begin{aligned} P(x^k) &\triangleq \prod_{j=1}^k P(x_j) \\ Q_*(y^k) &\triangleq \prod_{j=1}^k Q_*(y_j) \\ \mathbb{P}(x^{n-k}, y^{n-k}) &\triangleq \prod_{j=1}^k P(x_j) Q(y_j | x_j). \end{aligned}$$

Further, using Fact 2

$$\begin{aligned} \sum_{(x^k, y^k): \hat{P}_{x^k, y^k} = J_1} P(x^k) P_*(y^k) \\ = \sum_{x^k: \hat{P}_{x^k} = J_{1, \mathcal{X}}} P(x^k) \sum_{y^k: \hat{P}_{y^k} = J_{1, \mathcal{Y}}} Q_*(y^k) \\ \leq e^{-k(D(J_{1, \mathcal{X}} || P) + D(J_{1, \mathcal{Y}} || Q_*))} \\ \leq e^{-kD(J_{1, \mathcal{Y}} || Q_*)} \quad (94) \end{aligned}$$

where $J_{1, \mathcal{X}}$ and $J_{1, \mathcal{Y}}$ denote the left and right marginals of J , respectively, and where the second inequality follows by non-negativity of divergence.

A similar calculation yields

$$\begin{aligned} \sum_{(x^{n-k}, y^{n-k}): \hat{P}_{x^{n-k}, y^{n-k}} = J_2} \mathbb{P}(x^{n-k}, y^{n-k}) \\ \leq e^{-(n-k)D(J_2 || PQ)} \quad (95) \end{aligned}$$

From (93), (94), (95) and Fact 1 we get

$$\begin{aligned} \mathbb{P}_1(\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} = J) \\ \leq \text{poly}(n) \\ \times \max_{\substack{J_1 \in \mathcal{P}_k^{\mathcal{X}, \mathcal{Y}} \\ J_2 \in \mathcal{P}_{n-k}^{\mathcal{X}, \mathcal{Y}} \\ kJ_1 + (n-k)J_2 = nJ \\ k: \lceil n^{3/4} \rceil - 1 \leq k \leq n-1}} \exp \left[-k(D(J_{1, \mathcal{Y}} || Q_*) \right. \\ \left. - (n-k)D(J_2 || PQ) \right]. \quad (96) \end{aligned}$$

The maximum on the right-hand side of (96) is equal to

$$\begin{aligned} \max_{\substack{J_1 \in \mathcal{P}_k^{\mathcal{Y}} \\ J_2 \in \mathcal{P}_{n-k}^{\mathcal{Y}} \\ kJ_1 + (n-k)J_2 = nJ_{\mathcal{Y}} \\ k: \lceil n^{3/4} \rceil - 1 \leq k \leq n-1}} \exp \left[-kD(J_1 || Q_*) \right. \\ \left. - (n-k)D(J_2 || (PQ)_{\mathcal{Y}}) \right]. \quad (97) \end{aligned}$$

We upper bound the argument of the above exponential via the log-sum inequality to get

$$\begin{aligned} -kD(J_1 || Q_*) - (n-k)D(J_2 || (PQ)_{\mathcal{Y}}) \\ \leq -nD(J_{\mathcal{Y}} || \delta Q_* + (1-\delta)(PQ)_{\mathcal{Y}}), \quad (98) \end{aligned}$$

where $\delta \triangleq k/n$. Using (98), we upper-bound expression (97) by

$$\begin{aligned} \max_{\delta: n^{-1/4} - n^{-1} \leq \delta \leq 1} \exp \left[-nD(J_{\mathcal{Y}} || \delta Q_* + (1-\delta)(PQ)_{\mathcal{Y}}) \right] \\ \leq \max_{\delta: n^{-1/4} - n^{-1} \leq \delta \leq 1} \exp \left[-n\Omega(\delta^2) \right] \\ \leq \exp \left[-\Omega(n^{1/2}) \right], \quad (99) \end{aligned}$$

where for the first inequality we used Pinsker's inequality [7, Problem 17 p. 58]

$$D(P_1 || P_2) \geq \frac{1}{2 \ln 2} \|P_1 - P_2\|^2,$$

and assume that μ is small enough and n is large enough for this inequality to be valid. Such μ and n exist whenever the distributions Q_* and $(PQ)_{\mathcal{Y}}$ are different.

It then follows from (96) that

$$\mathbb{P}_1(\hat{P}_{C^n(1), Y_{\nu+i-n+1}^{\nu+i}} = J) \leq \exp \left[-\Omega(n^{1/2}) \right],$$

hence, from (91) and Fact 1 we get

$$\mathbb{P}_1(\tau_n = \nu + i) \leq \exp \left[-\Omega(n^{1/2}) \right]$$

for $i \in \{0, 1, \dots, g(n)\}$. Finally a union bound over times yields the desired result (89) since $g(n) = O(n)$.

APPENDIX B PROOF OF THEOREM 5

The desired Theorem is a stronger version of [7, Corollary 1.9, p. 107], and its proof closely follows the proof of the latter.

Before proceeding, we recall the definitions of η -image and l -neighborhood of a set of sequences.

Definition 4 (η -image, [7] Definition 2.1.2 p. 101): A set $\mathcal{B} \subset \mathcal{Y}^n$ is an η -image of a set $\mathcal{A} \subset \mathcal{X}^n$ if $Q(\mathcal{B} | x) \geq \eta$ for all $x \in \mathcal{A}$. The minimum cardinality of η -images of \mathcal{A} is denoted $g_Q(\mathcal{A}, \eta)$.

Definition 5 (l -neighborhood, [7] p. 86): The l -neighborhood of a set $\mathcal{B} \subset \mathcal{Y}^n$ is the set

$$\Gamma^l \mathcal{B} \triangleq \{y^n \in \mathcal{Y}^n : d_H(\{y^n\}, \mathcal{B}) \leq l\}$$

where $d_H(\{y^n\}, \mathcal{B})$ denotes the Hamming distance between y^n and \mathcal{B} , i.e.,

$$d_H(\{y^n\}, \mathcal{B}) = \min_{\tilde{y}^n \in \mathcal{B}} d_H(y^n, \tilde{y}^n).$$

As other notation, for a given conditional probability $Q(y|x)$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and $x^n \in \mathcal{X}^n$, we define the set

$$\mathcal{T}_{[Q]}^n(x^n) = \left\{ y^n \in \mathcal{Y}^n : \left| \hat{P}_{x^n, y^n}(a, b) - \hat{P}_{x^n}(a)Q(b|a) \right| > q, \forall (a, b) \in \mathcal{X} \times \mathcal{Y} \right\}$$

for a constant $q > 0$. To establish Theorem 5, we make use of the following three lemmas. Since we restrict attention to block coding schemes, i.e., coding scheme whose decoding happens at the fixed time n , we denote them simply by (\mathcal{C}_n, ϕ_n) instead of $(\mathcal{C}_n, (\gamma_n, \phi_n))$.

In the following, ϵ_n is always given by

$$\epsilon_n = (n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|} \exp(-nq^2/(2 \ln 2)).$$

Lemma 2: Given $\gamma \in (0, 1)$, $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, $P \in \mathcal{P}_n^{\mathcal{X}}$, and $\mathcal{A} \subset \mathcal{T}_P^n$, there exist (\mathcal{C}_n, ϕ_n) for each $n \geq n_0(\gamma, q, |\mathcal{X}|, |\mathcal{Y}|)$ such that

- 1) $c^n(m) \in \mathcal{A}$, for all $c^n(m) \in \mathcal{C}_n$
- 2) $\phi_n^{-1}(m) \subset \mathcal{T}_{[Q]}^n(c^n(m))$, $m \in \{1, 2, \dots, M\}$
- 3) the maximum error probability is upper bounded by $2\epsilon_n$
- 4) the rate satisfies

$$\frac{1}{n} \ln |\mathcal{C}_n| \geq \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon_n) - H(Q|P) - \gamma.$$

Proof of Lemma 2: The proof closely follows the proof of [7, Lemma 1.3, p. 101] since it essentially suffices to replace ϵ and γ in the proof of [7, Lemma 1.3, p. 101] with $2\epsilon_n$ and ϵ_n , respectively. We therefore omit the details here.

One of the steps of the proof consists in showing that

$$Q(\mathcal{T}_{[Q]}^n(x^n)|x^n) \geq 1 - \epsilon_n \quad (100)$$

for all $x^n \in \mathcal{X}^n$. To establish this, one proceeds as follow. Given $P \in \mathcal{P}_n^{\mathcal{X}}$ let \mathcal{D} denote the set of empirical conditional distributions $W(y|x) \in \mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}$ such that

$$|\hat{P}_{x^n}(a)W(b|a) - \hat{P}_{x^n}(a)Q(b|a)| > q$$

for all $(a, b) \in \mathcal{X} \times \mathcal{Y}$. We have

$$1 - Q(\mathcal{T}_{[Q]}^n(x^n)|x^n) = \sum_{W \in \mathcal{D} \cap \mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}} Q(\mathcal{T}_W^n(x^n)|x^n) \quad (101)$$

$$\leq \sum_{W \in \mathcal{D} \cap \mathcal{P}_n^{\mathcal{Y}|\mathcal{X}}} e^{-nD(W||Q|P)} \quad (102)$$

$$\leq (n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|} \exp(-n \min_{W \in \mathcal{D}} D(W||Q|P)) \quad (103)$$

$$\leq (n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|} \exp(-n \min_{W \in \mathcal{D}} \|PW - PQ\|^2 / 2 \ln 2) \quad (104)$$

$$\leq (n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|} \exp(-nq^2/(2 \ln 2)) \quad (105)$$

$$= \epsilon_n,$$

which shows (100). Inequality (102) follows from Fact 3, (103) follows from Fact 1, (104) follows from Pinsker's inequality

(see, e.g., [7, Problem 17, p. 58]), and (105) follows from the definition of \mathcal{D} . ■

Lemma 3 ([7, Lemma 1.4, p. 104]): For every $\epsilon, \gamma \in (0, 1)$, if (\mathcal{C}_n, ϕ_n) achieves an error probability ϵ and $\mathcal{C}_n \subset \mathcal{T}_P^n$, then

$$\frac{1}{n} \ln |\mathcal{C}_n| < \frac{1}{n} \ln g_Q(\mathcal{C}_n, \epsilon + \gamma) - H(Q|P) + \gamma$$

whenever $n \geq n_0(|\mathcal{X}|, |\mathcal{Y}|, \gamma)$.

Since this lemma is established in [7, Lemma 1.4, p. 104], we omit its proof.

Lemma 4: For every $\gamma > 0$, $\epsilon \in (0, 1)$, $Q \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, and $\mathcal{A} \subset \mathcal{X}^n$

$$\left| \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon) - \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon_n) \right| < \gamma$$

whenever $n \geq n_0(\gamma, q, |\mathcal{X}|, |\mathcal{Y}|)$.

Proof of Lemma 4: By the Blowing Up Lemma [7, Lemma 1.5.4, p. 92] and [7, Lemma 1.5.1, p. 86], given the sequence $\{\epsilon_n\}_{n \geq 1}$, there exist $\{l_n\}$ and $\{\eta_n\}$ such that $l_n/n \xrightarrow{n \rightarrow \infty} 0$ and $\eta_n \xrightarrow{n \rightarrow \infty} 1$, and such that the following two properties hold.

For any $\gamma > 0$ and $n \geq n_0(\gamma, q, |\mathcal{X}|, |\mathcal{Y}|)$

$$\frac{1}{n} \ln |\Gamma^{l_n} \mathcal{B}| - \frac{1}{n} \ln |\mathcal{B}| < \gamma \quad \text{for every } \mathcal{B} \subset \mathcal{Y}^n, \quad (106)$$

and for all $x^n \in \mathcal{X}^n$,

$$Q(\Gamma^{l_n} \mathcal{B}|x^n) \geq \eta_n \quad \text{whenever } Q(\mathcal{B}|x^n) \geq \epsilon_n. \quad (107)$$

Now, assuming that \mathcal{B} is an ϵ_n -image of \mathcal{A} with $|\mathcal{B}| = g_Q(\mathcal{A}, \epsilon_n)$, the relation (107) means that $\Gamma^{l_n} \mathcal{B}$ is an η_n -image of \mathcal{A} . Therefore we get

$$\begin{aligned} \frac{1}{n} \ln g_Q(\mathcal{A}, \eta_n) &\leq \frac{1}{n} \ln |\Gamma^{l_n} \mathcal{B}| \\ &\leq \gamma + \frac{1}{n} \ln |\mathcal{B}| \\ &= \gamma + \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon_n) \end{aligned} \quad (108)$$

where the second inequality follows from (106). Finally, since $\eta_n \rightarrow 1$ and $\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$, for n large enough we have

$$g_Q(\mathcal{A}, \epsilon) \leq g_Q(\mathcal{A}, \eta_n) \quad \text{and} \quad \epsilon_n \leq \epsilon,$$

and therefore from (108) we get

$$\frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon) \leq \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon_n) \leq \gamma + \frac{1}{n} \ln g_Q(\mathcal{A}, \epsilon)$$

yielding the desired result. ■

We now use these lemmas to establish Theorem 5. Choose $\epsilon, \gamma > 0$ such that $\epsilon + \gamma < l$. Let (\mathcal{C}_n, ϕ_n) be a coding scheme that achieves maximum error probability ϵ . Without loss of generality, we assume that $\mathcal{C}_n \subset \mathcal{T}_P^n$ (If not, group codewords into families of common type. The largest family of codewords has error probability no larger than ϵ , and its rate is essentially the same as the rate of the original code \mathcal{C}_n .) Therefore

$$\begin{aligned} \frac{1}{n} \ln |\mathcal{C}_n| &\leq \frac{1}{n} \ln g_Q(\mathcal{C}_n, \epsilon + \gamma) - H(Q|P) + \gamma \\ &\leq \frac{1}{n} \ln g_Q(\mathcal{C}_n, l) - H(Q|P) + \gamma \\ &\leq \frac{1}{n} \ln g_Q(\mathcal{C}_n, \epsilon_n) - H(Q|P) + 2\gamma \end{aligned} \quad (109)$$

for $n \geq n_o(\gamma, l, |\mathcal{X}|, |\mathcal{Y}|)$, where the first and third inequalities follow from Lemmas 3 and 4, respectively, and where the second inequality follows since $g_Q(\mathcal{C}_n, \epsilon)$ is nondecreasing in ϵ . On the other hand, by Lemma 2, there exists a coding scheme $(\mathcal{C}'_n, \phi'_n)$, with $\mathcal{C}'_n \subset \mathcal{C}_n$ that achieves a probability of error upper bounded by $2\epsilon_n$ and such that its rate satisfies

$$\frac{1}{n} \ln |\mathcal{C}'_n| \geq \frac{1}{n} \ln g_Q(\mathcal{C}_n, \epsilon_n) - H(Q|P) - \gamma \quad (110)$$

for $n \geq n_o(\gamma, q, |\mathcal{X}|, |\mathcal{Y}|)$. From (109) and (110) we deduce the rate of \mathcal{C}'_n is lower bounded as

$$\frac{1}{n} \ln |\mathcal{C}'_n| \geq \frac{1}{n} \ln |\mathcal{C}_n| - 3\gamma$$

whenever $n \geq n_o(\gamma, l, q, |\mathcal{X}|, |\mathcal{Y}|)$. This yields the desired result. ■

REFERENCES

- [1] A. Tchamkerten, V. Chandar, and G. Wornell, "On the capacity region of asynchronous channels," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2008.
- [2] V. Chandar, A. Tchamkerten, and G. Wornell, "Training-based schemes are suboptimal for high rate asynchronous communication," in *Proc. IEEE Information Theory Work. (ITW)*, Taormina, October 2009.
- [3] A. Tchamkerten, V. Chandar, and G. Wornell, "Communication under strong asynchronism," *IEEE Trans. Inform. Th.*, vol. 55, no. 10, pp. 4508–4528, October 2009.
- [4] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423, October 1948.
- [5] V. Chandar, A. Tchamkerten, and D. Tse, "Asynchronous capacity per unit cost," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, june 2010, pp. 280–284.
- [6] —, "Asynchronous capacity per unit cost," *CoRR*, vol. abs/1007.4872, 2010.
- [7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels*. New York: Academic Press, 1981.
- [8] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, 2nd edition*. MIT Press, McGraw-Hill Book Company, 2000.
- [9] V. Chandar, A. Tchamkerten, and G. Wornell, "Optimal sequential frame synchronization," *IEEE Trans. Inform. Th.*, vol. 54, no. 8, pp. 3725–3728, 2008.
- [10] I. Csiszár and P. Narayan, "Arbitrarily varying channels with constrained inputs and states," *IEEE Transactions on Information Theory*, vol. 34, no. 1, pp. 27–34, 1988.
- [11] T. Cover and J. Thomas, *Elements of information theory*. New York: Wiley, 2006.
- [12] R. G. Gallager, *Information Theory and Reliable Communication*. Budapest: Wiley, 1968.
- [13] D. Wang, V. Chandar, S.-Y. Chung, and G. W. Wornell, "Error exponents in asynchronous communication," in *Proc. Int. Symp. Inform. Theory*, St. Petersburg, Russia, July 2011.

